TESIS DOCTORAL

2023

Metodología para la construcción Índices de Precios de la Vivienda robustos, segmentados y a escala, basados en modelos de valoración hedónica

DAVID REY BLANCO

PROGRAMA DE DOCTORADO EN ECONOMÍA Y EMPRESA

DIRECTOR: Dr. JULIO GONZÁLEZ ARIAS

Agradecimientos

En primer lugar, deseo reconocer y expresar mi agradecimiento al Dr. Julio González Arias, director de tesis, por su inestimable orientación, apoyo y dedicación en el desarrollo de este trabajo. Sus conocimientos, experiencia y compromiso han sido fundamentales para mi formación.

Quiero también expresar mi más profundo reconocimiento al Dr. Juan Vicente, tristemente desaparecido, cuya dedicación y experiencia han sido fundamentales para el desarrollo de esta investigación doctoral, y no existen palabras suficientes para expresar mi gratitud y respeto hacia su labor académica.

También quiero agradecer a Nerea, Álvaro y Daniela, quienes han sido una fuente constante de apoyo e inspiración a lo largo de este proceso. Destacar que sus sacrificios y comprensión por todo el tiempo que os he robado, y que han permitido dedicar esfuerzo a esta investigación, y por eso les estoy infinitamente agradecido.

Aprovecho este espacio para expresar mi más sincero agradecimiento a mis padres, quienes sembraron en mí desde temprana edad el anhelo por el conocimiento. Su constante respaldo, amor y sacrificio incondicional durante mis años de formación, me han permitido llegar hasta aquí. Esta investigación y muchos de mis logros profesionales son un reflejo de vuestra dedicación.

Además, quisiera extender mi gratitud hacia mis compañeros de trabajo en Idealista, quienes han compartido sus ideas, conocimientos y experiencias en el ámbito académico. Sus valiosas aportaciones han enriquecido mi trabajo y me han impulsado a superar obstáculos y desafíos en este camino.

Por último, doy las gracias a mis amigos y seres queridos por su apoyo y palabras de aliento a lo largo de este viaje, que han sido esenciales para superar los momentos de flaqueza.

A mi familia

Resumen

El sector inmobiliario español posterior a la gran crisis financiera de finales de la primera década de los 2000, dio lugar a un mercado mucho más orientado hacia el alquiler. La escasez de oferta junto con una creciente demanda produjo un desequilibrio debido a una oferta estática (la crisis también frenó completamente la nueva construcción) y una demanda en auge. Lo cual presionó los precios de forma ascendente hasta la actualidad. Además, durante casi una década de restricción crediticia, los colectivos con menos recursos económicos han sido los más afectados por este fenómeno. El aumento de las rentas ha atraído una intensa atención social, que, consecuentemente, ha suscitado el interés de la administración, poniendo en práctica distintas acciones. Sin embargo, las mismas requieren un conocimiento actualizado y profundo de la realidad inmobiliaria, y precisan de fuentes de información acordes. Desafortunadamente, no existe un registro completo nacional y actualizado de información del alquiler, y la única base de datos de esta naturaleza, publicada por la administración tiene un nivel de detalle limitado y ofrece información con años de retraso.

El uso generalizado de los portales inmobiliarios en el proceso de búsqueda, y su alta correlación con los datos reales, hacen de los ellos un instrumento ideal para la creación de un índice de precios de la vivienda con las características deseables: fidelidad con el dato real, representatividad, nivel de detalle y actualización inmediata. No obstante, esta fuente presenta retos importantes derivados de la calidad de la información y de los numerosos sesgos a controlar.

La presente metodología desarrolla de forma eficaz un índice de la vivenda a partir de datos del portal Idealista, que junto con fuentes públicas catastrales y del INE logra las características deseadas. Dicho índice se presenta como una herramienta adecuada para la toma de decisiones y el control continuo del mercado, así como un instrumento de análisis con potencial para la estimación de movimientos futuros el corto plazo.

Palabras clave: *Inmobiliaria, Índice de Precios de la Vivienda, Alquiler, Modelo Hedónico.*

Abstract

The Spanish residential real estate sector headed to a market much more oriented towards rental after the major financial crisis in the late 2000s. The scarcity of supply, coupled with a growing demand, gave rise to an imbalanced market due to static supply (the crisis also completely halted new construction) and booming demand. This resulted in upward pressure on prices still persists today. Furthermore, during almost a decade of credit restriction, economically disadvantaged groups have been the most affected ones by this phenomenon.

The increase in rents has lured significant social attention, which, consequently, has sparked the interest of the government, leading to the implementation of various actions. However, these actions require up-to-date and in-depth knowledge of the real estate reality, and demand suitable sources of information. Unfortunately, there is no comprehensive national and up-to-date rental information registry, and the only database of this nature published by the government has limited detail and provides information with years of delay.

The widespread use of real estate portals in the search process, and their high correlation with actual data, make them an ideal tool for creating house price indices with the desirable characteristics: fidelity to real data, representativeness, level of detail, and immediate updates. However, this source presents significant challenges due to the quality of the information and the numerous biases that need to be controlled for.

The present methodology effectively develops a house price index using data from the Idealista portal, which, along with public sources from the Cadastre and the National Institute of Statistics, achieves the desired characteristics. This index is presented as a suitable tool for decision-making and continuous market monitoring, as well as an analytical instrument with potential for estimating short-term future movements.

Key words: *Real Estate, House Price Index, Rental, Hedonic Model.*

Índice

Índice de Tablas

Índice de Figuras

Glosario

Acrónimo	Definición
API	El Agente de la Propiedad Inmobiliaria (API) es un profesional cualificado que opera en el mercado inmobiliario y que está regulado por Real Decreto 1294/2007
AVM	Modelo de valoración automática (de la vivienda)
CAM	Comunidad Autónoma de Madrid
CBD	Central Business District, se refiere al distrito central de la ciudad, donde se desarrolla principalmente la actividad comercial
COD	Coeficiente de dispersión (estadística)
DGGS	Sistema discreto global en malla, sistema de referencia geográfica
DIW	Deutsches Institut für Wirtschaftsforschung, Instituto Alemán de Estudios Económicos
EPF	Encuesta de Presupuestos Familiares
EU-SILC	Estadística europea de ingresos y condiciones de vida
GAM	Modelo Aditivo Generalizado
GREG	Estimador General de Regresión
HPI	House Price Index
INCASOL	Instituto Catalán del Suelo
INE	Instituto Nacional de Estadística
IPC	Índice de Precios al Consumo
IPV	Índices de Precios de la Vivienda
IRPF	Impuesto sobre la Renta de las Personas Físicas
MAE	Error Medio Absoluto
MAPE	Error Medio Absoluto en porcentaje
MCO	(Regresión) por mínimos cuadrados

MEDAE	Error Mediano Absoluto
MEDAPE	Error Mediano Absoluto en porcentaje
MITMA	Ministerio de Transportes, Movilidad y Agenda urbana
MLE	Estimador de Máxima Verosimilitud
OCDE	Organización para la Cooperación y el Desarrollo Económicos
OLS	Regresión por Mínimos Cuadrados (en inglés)
OOB	Out Of Bag sample, muestra de control fuera del aprendizaje del modelo
OOHPI	Owner Occupied House Price Index
OSM	Open Street Map
P2P	P2P del original en inglés peer to peer, se refiere a un mercado no intermediado en el que los particulares intercambian servicios directamente entre ellos
PCA	Análisis de Componentes Principales
POI	Punto de interés
RMSE	Raíz cuadrada del error medio cuadrático medio
SVM	Máquina de soporte vectorial, algoritmo de aprendizaje estadístico
UE	Unión Europea

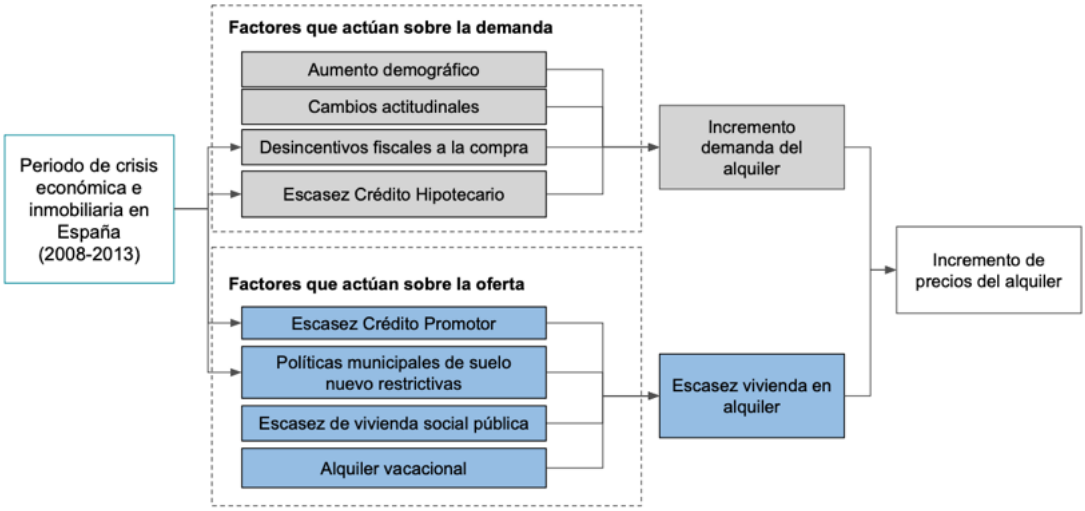
Introducción

> *“Los inmuebles no se pueden perder, robar y nadie puede llevárselos. Adquiridos con sentido común, pagados por completo y gestionados con razonable cuidado, son la inversión más segura del mundo.”*
>
> — Franklin D. Roosevelt

Motivación

El sector inmobiliario español que sucedió a la gran crisis financiera de finales de la primera década de los 2000, dio lugar a un mercado residencial con mucho más peso del alquiler. Una limitada oferta junto con demanda cada vez más pujante produjo un desequilibrio del mercado, con una oferta que no podía acompasar la demanda (la crisis también frenó completamente la nueva construcción). Lo cual, unido a otros factores que alimentaban la restricción de oferta y el crecimiento de la demanda, presionó los precios de forma ascendente hasta la actualidad (Figura 1). Además, durante casi una década de restricción crediticia, los colectivos con menos recursos económicos han sido los más afectados por este fenómeno (Banco de España, 2021). El aumento de las rentas ha atraído una intensa atención social, que, consecuentemente, ha suscitado el interés de la administración, que ha puesto en práctica distintas regulaciones que precisan de un conocimiento profundo y actualizado de la realidad inmobiliaria. Lo cual ha puesto de manifiesto la necesidad de fuentes de información adecuadas para esta finalidad.

Figura 1. Factores del mercado clave del mercado inmobiliario del alquiler español en el periodo 2007-actualidad



Fuente: elaboración propia.

Históricamente, el estudio de la evolución de los precios de la vivienda se ha centrado en índices de precio de compraventa. En buena medida porque apenas existen registros públicos que reflejen adecuadamente la evolución del precio de alquiler, bien porque su cobertura sea muy local o porque tiene debilidades metodológicas. Incluso cuando estos datos existen, un retraso de meses en la información pueden suponer una importante distorsión de la percepción del mercado de sus agentes principales (Anenberg y Laufer, 2017). El impacto de esta cuestión es crucial, puesto que a lo largo de las dos últimas décadas la proporción de personas que viven en alquiler con respecto al régimen en propiedad es cada vez mayor, tendencia que confluye con la situación que se observa en la gran mayoría de los países europeos de nuestro entorno (Eurostat, 2022). Un catalizador de este fenómeno fue el inicio de la última gran crisis inmobiliaria en España, a finales de la segunda mitad de la primera década del siglo XXI. Como consecuencia, los principales agentes del mercado inmobiliario y la administración han incrementado su interés en disponer de más información detallada y actualizada de la actividad del mercado inmobiliario del alquiler, aunque para ello es necesario activar nuevas líneas de investigación en este campo. Un primer punto de partida para cubrir los intereses previos es, al menos, disponer de la evolución de precios desglosada por distritos, barrios y municipios de manera mensual, y para cada uno de los distintos tipos de inmueble. Además, esta información debe basarse en el dato más reciente posible, para ofrecer solvencia y robustez a los análisis que la utilizan.

La evolución que ha experimentado el mercado inmobiliario en España en los

últimos veinte años ha influido de forma notable en todo el sistema económico, y con toda probabilidad, su evolución futura condicionará las decisiones a tomar en muchos sectores, particularmente en el sistema financiero. No se trata de una cuestión endémica de la economía española, ya que como indican Case y Quigley (2008), los efectos de la caída de los precios de las viviendas afectan a los mercados financieros mundiales e impactan directamente en la riqueza de los agentes económicos. En el caso particular español, podemos constatar cómo los ciclos de expansión del mercado de la construcción e inmobiliario de inicios de los 2000, dieron lugar a una economía altamente dependiente del este sector de la economía y enormemente apalancada, que se tradujo en una gran crisis económica. Además, el alto endeudamiento de los agentes económicos y de las familias, junto con la alta exposición a la deuda hipotecaria de los bancos, supuso una restricción al crédito que provocó importantes cambios en el mercado de la vivienda y reforzó la importancia de la vivienda en alquiler. Este impulso en la demanda, junto con un insuficiente movimiento en la oferta de inmuebles de alquiler, provocó un efecto inflacionista en los precios que aún observamos en la actualidad. Como es lógico, han sido los colectivos más desfavorecidos, los que no tienen más opción habitacional que el alquiler, los principales perjudicados socialmente por fenómeno.

El sector de la construcción está íntimamente ligado al inmobiliario y actúa de forma fundamental en la formación de la oferta. Este área de actividad fue uno de los motores económicos en la primera mitad de la década de los 2000 con una importante participación en el PIB español. Esta cuota alcanzó el máximo en 2006 con un 10,4% y cayó consecuencia de la crisis financiera e inmobiliaria en España, decreciendo en los siguientes años hasta alcanzar un punto de inflexión en el año 2014, con un 5,1% de participación sobre el PIB. A partir de este año, se observaron ligeras subidas interanuales hasta alcanzar el 5,8 % en el año 2019[1]. En el caso de la construcción de viviendas residenciales, la tendencia creciente de los primeros años del siglo XXI se invirtió en 2008, cuando el número de licencias residenciales cayó un 42,3% y que continuó con caídas superiores al 50 % hasta el año 2012. En 2013 se invirtió la tendencia con un incremento del 7 %, que ha continuado creciendo hasta la actualidad, tal como recoge la información del Ministerio de Transportes[2].

A su vez, la generación de un gran stock de viviendas en manos del sector bancario desde 2008, principalmente representada por la Sociedad de Gestión de Activos Procedentes de la Re-estructuración Bancaria[3] (en adelante SAREB), se esperaba

[1] INE (Cuentas Económicas).

[2] Ministerio de Transportes, Movilidad y Agenda Urbana (Datos de licencias municipales de obra - Número de viviendas según obra).

[3] A la SAREB le fueron transferidos una serie de activos inmobiliarios denominados tóxicos y

que tuviera un impacto en el *stock* disponible de vivienda en alquiler de forma que pudiera haber actuado como estabilizador de precios, lo cual no se produjo.

SAREB se creó en 2012 como un mecanismo para hacer frente a los problemas del sector bancario: saneamiento de balances; aumento de requerimientos mínimos de capital, reestructuración del sector de las cajas de ahorro y el incremento de forma significativa de los requisitos relativos a provisiones por préstamos con promociones inmobiliarias. SAREB es mayoritariamente de capital privado, aunque un 45 % del capital pertenece al Fondo de Reestructuración Ordenada Bancaria (en adelante FROB), que es de titularidad pública. La vida prevista para esta sociedad se estableció en 15 años, en los que estaba previsto que la sociedad se desprendiera de los activos problemáticos. La entidad tenía en cartera, al cierre del ejercicio de 2019[a], activos de difícil colocación por un valor superior a 32.000 millones de euros, repartidos en: 19.400 millones en préstamos (aproximadamente la mitad del volumen en su creación), un total de 12.800 millones en inmuebles, procedentes de la cartera heredada de la banca cuyo valor era 50.781 millones de euros. El ritmo de drenaje del volumen de activos tóxicos desde el inicio hace difícil lograr el objetivo de completarse en 2027, y por tanto, su papel esperado como estabilizador de precios y como medio de control del número de nuevos desarrollos es cuestionable. En particular, si atendemos al estudio de impacto socio-económico publicado en 2019 por la propia SAREB[b], donde es innegable que la contribución económica, financiera y tributaria valorada en sus primeros seis años de vida, y valorada en 27.329 millones de euros es importante, pero no lo es tanto su impacto en el sector inmobiliario residencial con un total de 89.500 inmuebles vendidos y 7.552 viviendas alquiladas, que en términos anuales resultan aproximadamente 15.000 y 1.260 viviendas al año respectivamente, cifras que para el caso de compraventa anual representaría un 2.5% del total de compraventas en el año 2018[c], para el caso del alquiler el volumen anual representa un 0,029% del total de hogares españoles en régimen de alquiler para el año 2018[d].

[a] SAREB (Informe de actividad 2020).
[b] SAREB (Estudio de impacto socioeconómico, actualización 2018, Marzo 2019).
[c] Estadísticas del Banco de España, Indicadores del Mercado de la Vivienda (23 de septiembre de 2021).
[d] Eurostat (18,63 millones de hogares en 2018) e INE (tasa de tenencia en alquiler del 23%).

La política de la vivienda no logrado atender completamente a las causas raíces de los problemas del mercado del alquiler (Arruñada, 2022), y las normativas que

procedentes de cinco entidades bancarias nacionalizadas: BFA-Bankia, Catalunya Banc, Novagalicia Banco, Banco Gallego y Banco de Valencia, y entidades en proceso de reestructuración o resolución: Banco Mare Nostrum, CEISS, Caja 3 y Liberbank.

se han implementado hasta el momento no han conseguido erradicar, entre otras cuestiones, la escasez de oferta. Alguna de estas medidas, como las centradas en establecer un máximo a los precios o sus incrementos anuales, no han logrado el propósito de reducir los precios del alquiler en el medio plazo. El control de rentas se ha estudiado en profundidad desde hace más de medio siglo, desde un punto de vista teórico (Bloomberg, 1947; Smith y Tomlinson, 1981), o con evidencias empíricas, como el del experimento natural en San Francisco en (Diamond *et al.*, 2019). En España, estas medidas se han articulado, en la mayor parte de los casos, sobre la limitación de los precios máximos del alquiler en aquellas zonas con mayor tensión entre oferta y demanda (Monràs y Montalvo, 2022), o bien estableciendo un incremento máximo en la actualización de las rentas anuales.

La asimetría de información de mercado entre oferentes y demandantes, tanto en segmento residencial como en comercial (Arruñada, 2022), provoca desequilibrios en los precios y falta de transparencia ante los ciudadanos acerca de qué factores que causan las alzas de precios (Lacerda, 2018). La cual, desde un ángulo micro, se sustancia en la desconfianza entre agentes (arrendador sobre el arrendatario), y que se ve agravada por una regulación que prima los alquileres a largo plazo. Todo ello da lugar a que el arrendador incluya una prima sobre el precio por la posibilidad de impago, y que lo haga sin tener en cuenta la solvencia real del inquilino. La ausencia de transparencia desde un punto de vista macro ligada a mejorar el funcionamiento del mercado, apenas se ha resuelto dada la escasez de información suficientemente objetiva, actualizada, y estudios exhaustivos al respecto.

España es también uno de los destinos turísticos más importantes del mundo. A pesar que el los turistas se alojan principalmente en hoteles el uso del alquiler vacacional, es cada vez más habitual, con un incremento total de 9,7 millones de personas en 2018 (INE, 2023a). Este factor presiona de forma adicional a la oferta, ya que parte de los inmuebles disponibles se ponen en régimen de alquiler por estancias cortas o noches por su mayor rentabilidad en el corto plazo (Ayouba *et al.*, 2020; Franco y Santos, 2021; Koster *et al.*, 2018). Este fenómeno no es aplicable a todos los mercados sino a los lugares de mayor interés turístico, que en algunos casos como Sevilla, Madrid, Málaga o Barcelona coinciden con el área metropolitana donde ya existe un gran nivel de demanda del alquiler de largo plazo. Lo cual ha dado lugar a un amplio campo de estudio, no solamente en España, con estudios en diversas partes del mundo: Barcelona (Garcia-López *et al.*, 2020); para Lyon, Montpellier y París (Francia) (Ayouba *et al.*, 2020); en los Estados Unidos (Koster *et al.*, 2018), (Barron *et al.*, 2021) y (Horn y Merante, 2017); y en varias ciudades de Portugal (Franco y Santos, 2021).

El impacto del mercado del alquiler en la economía ha ido en aumento por el creciente peso del número de hogares cuyo régimen de tenencia es de este tipo, frente a aquellos en propiedad. Esta relación partía de un 21,5 % de hogares en alquiler en 2004 hasta el 23,3 % en 2017[4]; dicha tendencia se intensificó a partir del año 2013 para las viviendas a precio de mercado, mientras que en los segmentos de alquiler social se redujeron en el periodo 2005-2018 (Banco de España, 2019). La tensión en vivienda social es preocupante, tal y como indica el Banco de España (Román *et al.*, 2020), por la mayor exposición a estos incrementos de precios de los colectivos más desfavorecidos de la sociedad, como aquellos que tienen rentas más bajas o los jóvenes.

Así pues, desde 2020 en adelante el mercado del alquiler se encuentra en una situación sin precedentes, en cuanto a su desarrollo y en relación al sistema de formación de precios. Los cambios bruscos en el mercado experimentados en la crisis del Covid-19, que dieron lugar a cambios en el tipo de vivienda más buscada (más amplia o fuera del centro de las ciudades), han despertado el interés por información más inmediata y detallada (Biancotti *et al.*, 2020). De forma que se pueda conocer los efectos en el mercado inmobiliario de externalidades o cambios de ciclo económico, para las que no son suficiente los registros oficiales disponibles.

Desde el ángulo de la macroeconomía, la literatura ha tratado habitualmente la vivienda como un bien homogéneo, sobre el que no se tienen en consideración las características que hacen a cada vivienda un elemento singular. Esta singularidad se apoya en tres aspectos: heterogeneidad, durabilidad e inmovilidad (Kiel y Zabel, 2008), y por tanto, precisa herramientas de análisis que incorporen dicha diversidad, dada la exposición de los sistemas financieros a las variaciones de los precios de la vivienda (Anundsen *et al.*, 2016). En los modelos microeconómicos es común incorporar el carácter único de la vivienda a través de sus cualidades, con este objetivo los modelos hedónicos surgen de la idea de que los bienes son valorados por sus atributos portadores de utilidad (Rosen, 1974). En el mercado inmobiliario, los atributos pueden incluir características estructurales de la vivienda como metros cuadrados, número de dormitorios/baños, antigüedad, estado, etc., así como aspectos de ubicación como distritos escolares, proximidad a servicios, acceso al transporte y demografía del vecindario.

La ubicación de la vivienda forma parte de su singularidad y debe incorporarse en los modelos. La creación de variables que representan el aspecto geográfico comenzó los años 60 y 70 del siglo pasado con Alonso (1964), Mills (1972), Ball (1974; 1973) o Wilkinson (1974) modelando la influencia del vecindario

[4]INE (Hogares por régimen de tenencia de la vivienda - Encuesta de condiciones de vida).

en el precio. Esta cuestión comporta una serie de retos metodológicos: el control de la heterogeneidad y dependencia espacial (Anselin y Rey, 2014), tratar la autocorrelación espacial (Anselin y Griffith, 1988) y la reducción de la heterocedasticidad (Fletcher *et al.*, 2000). La incorporación de características de localización ofrece una serie de ventajas colaterales como: la identificación de los submercados de la vivienda (Goodman y Thibodeau, 1998), esenciales para analizar fenómenos de conflictividad social (Deng *et al.*, 2003; Quigley, 1995); o el análisis topológico de las ciudades (Henderson, 1985; Turnbull, 1990). Por otra parte, la cuestión se ha ligado al concepto de accesibilidad de servicios desde la ubicación de la vivienda (Batty, 2009), estableciendo el grado influencia de los precios residenciales con la cercanía de los puestos de trabajo, servicios de restauración y de ocio, o factores medioambientales de la zona (Azqueta Oyarzun, 1994; Palmquist, 1989).

El método de precios hedónicos utiliza el análisis de regresión para estimar el valor implícito de cada característica de la vivienda. Lo cual, generalmente, involucra a un gran número de variables con elevados índices de colinealidad, y afecta a las precisión de los métodos métodos lineales (Orford, 2017). Entre los diversos enfoques para solucionarlo, Kain y Quigley (1970) propusieron el uso de componentes principales para reducir el problema de la colinealidad. Por otra parte, la amplitud de parámetros tiene otras implicaciones negativas, como un menor nivel de interpretabilidad de la contribución de cada factor. Lo cual ha sido ampliamente estudiado e involucra la correcta especificación y tipo de algoritmo, de forma que produzca modelos económicamente interpretables (Berndt, 1991; Freeman, 1979; Guilkey *et al.*, 1989; Rosenthal, 1989).

La transformación digital del sector inmobiliario ha supuesto que la actividad de búsqueda de vivienda se traslade de forma casi total a la red. Dicho fenómeno se ha producido de manera progresiva desde aproximadamente el año 2000, en paralelo a la explosión de nuevos servicios y empresas de Internet en España. El 65% de los españoles eligen los portales como el mejor medio para encontrar casa y el el 74,6 % el porcentaje de ellos los que usan este medio para la búsqueda de vivienda, como indica la encuesta desarrollada por la consultora inmobiliaria Remax[5]. Estas cifras son similares al resto de países de la Unión Europea, donde de media un 81 % de los ciudadanos buscan vivienda en medios electrónicos.

El tránsito desde las promotoras, inmobiliarias, los Agentes de la Propiedad Inmobiliaria (en adelante APIs[6]) y otros intermediarios del canal presencial al digital ha originando cambios que afectan a todo el sector. Estos cambios afectan

[5]Encuesta “At Home in Europe”, informe: Most relevant search criteria. Fuente: Remax.

[6]El Agente de la Propiedad Inmobiliaria, API, es un profesional cualificado que opera en el mercado inmobiliario y que está regulado por Real Decreto 1294/2007.

desde la gestión de la relación con el demandante de vivienda a los importes de las comisiones pagadas sobre la operación. En este contexto, los portales no actúan como intermediarios en las operaciones de compraventa, sino que son un canal de comunicación de la oferta y tienen nula intervención en el establecimiento de los precios. Son los particulares y los profesionales inmobiliarios quienes interactúan directamente en proceso de venta en estas plataformas, que actúa como un mecanismo para agregar de forma conjunta a la oferta y la demanda.

En diciembre de 2022, según datos de Similarweb[7], se contabilizaron en España un total de 110 millones de páginas vistas en portales inmobiliarios, del que los cuatro portales principales: Idealista, Fotocasa, Habitaclia y Pisos.com, acapararon un 94 % del tráfico de búsqueda de viviendas. Este es un sector altamente concentrado, donde Idealista es líder con gran diferencia respecto al resto, con 43,5 millones de visitas mensuales (un 39,6 % del tráfico) con respecto a su inmediato competidor con 10,1 millones (un 10,8 % del tráfico). Aunque esto no ha sido así siempre, según de datos de Google Trends[8], la cuota de mercado de los dos líderes fue similar entre 2004 hasta 2015, consolidándose posteriormente Idealista como líder con amplia diferencia. En términos de contenido publicado, en septiembre de 2021, Fotocasa contaba con un total de 65.332 viviendas en alquiler y 710.375 en venta publicadas[9], mientras que Idealista contaba con 98.987 viviendas en alquiler y 684.073 en venta[10].

El uso de fuentes de portales para el cálculo de índices de la vivienda no es nuevo, una de las primeras referencias es un índice de compraventa y alquiler calculado sobre una base de datos de anuncios clasificados en el periódico “El Mercurio” de Santiago de Chile, sobre una serie de datos mensual de anuncios entre 1998 y 2002 (Desormeaux y Piguillem, 2003). Más recientemente, Anemberg y Laufer (2017) desarrollaron un índice de precios para la Reserva Federal de Estados Unidos basado en datos de oferta (de múltiples MLS en Estados Unidos). Para estos autores, el uso de esta fuente permite conocer las condiciones actuales del mercado, y además, esto índices resultantes comportamientos observados en los índices oficiales, dado el alto nivel de correlación entre la oferta y las transacciones reales (Kokot y Bas, 2015).

El efecto del retraso temporal entre la fecha de publicación de estadísticas del mercado y la fecha en la que se producen, genera una importante asimetría de información entre los distintos agentes del mercado inmobiliario que tiene efectos medibles en aspectos económicos clave. Por ejemplo, la publicación de los datos

[7] Similarweb (2022) es una compañía que ofrece servicios de análisis web, como medidas de volumen de tráfico y de usuarios a sitios de internet.
[8] Comparativa del interés en Idealista, Fotocasa, Habitaclia y Pisos.com por Google Trends.
[9] Fotocasa (datos en el portal, 12 de septiembre de 2021).
[10] Idealista (datos en el portal, 12 de septiembre de 2021).

de los índices de precio en Estados Unidos tienen un efecto inmediato en los mercados de valores de las compañías cotizadas, aunque esta información proceda de una situación de meses pasados (Anenberg y Laufer, 2017).

Para mitigar la ausencia de información actualizada, los bancos centrales, como los de Italia, Francia, España o Gran Bretaña, han trabajado en incorporar esta información. En el caso italiano, Loberto *et al.* (Loberto *et al.*, 2018) utiliza información del portal Inmobiliare.it, con frecuencia semanal, para construir un índice de la vivienda alternativo. En dicho caso se demuestra como el índice guarda una fuerte relación con la información del mercado, al concluir que los índices de oferta (portal) tienen una alta correlación con los índices basados en transacciones[11], con un R^2 de 0.96 entre ambas magnitudes. También evidencia ciertos problemas con la información, como que una misma propiedad que pueda estar anunciada más de una vez, o la cuestión de datos ausentes que se solventan usando el campo de texto abierto de comentarios para imputar esta información.

En el artículo de investigación del Banco de Francia, Bricogne *et al.* (2023) realizan un seguimiento de los precios diario de datos de cinco portales en el Reino Unido, además, mediante técnicas de aprendizaje automático desarrollan un modelo de correspondencia entre los precios de oferta y los registrados por los notarios. En Asia, Wang y Wu (2020) crean un índice de precios de la vivienda para 274 ciudades en China, utilizando datos de portales inmobiliarios.

Aunque las fuentes alternativas han dado lugar al diseño de índices inmobiliarios más allá del estudio de los precios, por ejemplo, Chauvet *et al.* (2013) y Alexander *et al.* (2014) construyeron índices basados en las búsquedas más habituales en Google[12], sobre el mercado inmobiliario y su regulación, donde se demuestra como es posible la construcción de un índice de "sentimiento" del mercado altamente correlacionado con los precios. Este tipo de análisis, ofrecen una gran de correlación con los resultados de índices tradicionales, como en el caso del índice de sentimiento de consumidores de la Universidad de Míchigan[13] (Bram y Ludvigson, 1998). Incluso otros artículos, como el de Galesi *et al.* (2020), estiman el poder de negociación, de propietarios y demandantes, en función del número de contactos de los anuncios en portales inmobiliarios.

Las fluctuaciones del precio de la vivienda tienen un impacto real en los ingresos, riqueza y capacidad de ahorro y bienestar social de los ciudadanos, como se ha constatado en las últimas fases del ciclo económico. De forma recíproca también las políticas monetarias condicionan los movimientos del mercado de la vivienda. Es un hecho, por ejemplo, que los periodos con tipos de interés bajos coincidan

[11] Datos estadísticos (OMI) basados en registros oficiales del Ministerio de Hacienda Italiano.
[12] A través de su servicio Google Trends.
[13] Información disponible en https://data.sca.isr.umich.edu.

con momentos de expansión del mercado inmobiliario residencial, en buena forma porque incita a la contratación de créditos hipotecarios. Por tanto, la vigilancia de los cambios en los precios de la vivienda son una tarea esencial para controlar la riqueza real de los individuos, y ayudan a calibrar los riesgos de estabilidad financiera derivados de los desequilibrios en los mercados de residenciales.

Los índices de precios de la vivienda (en adelante IPV) proporcionan un instrumento útil para entender las dinámicas del mercado de la vivienda (Case y Quigley, 2008), pero también para medir el estado de la economía real, la eficacia de la política fiscal del gobierno, el impacto de la política monetaria del banco central y el comportamiento de los mercados financieros. Cada vez más, los bancos centrales otorgan una importancia capital a los precios de los activos inmobiliarios residenciales y comerciales en el establecimiento de su política monetaria, con la esperanza de evitar los efectos desestabilizadores de sus auges y las caídas. Sin embargo, las características específicas de la propiedad residencial, al ser cada vivienda un elemento singular, dificultan la recopilación de datos primarios y el cálculo de índices. Esta cuestión se agrava de forma supranacional cuando la disparidad de datos en diferentes países y las características inmobiliarias específicas, dificultan disponer de criterios comunes entre países como apunta Diewert (2009).

El control de los precios desde un punto de vista macroeconómico se puede implementar de varias maneras, que van desde la fijación de precios de referencia de los activos, el establecimiento de objetivos de inflación de precios al consumidor, hasta modificar la forma en que se calcula el costo de los servicios de vivienda ocupada por el propietario en el Índice de Precios al Consumo (en adelante IPC). De manera más detallada Fenwick (2013) identifica cuatro usos potenciales de los IPV:

- Como indicador macroeconómico general.
- Como insumo en la medición de la inflación de precios al consumidor.
- Como elemento en el cálculo de la riqueza (real) de los hogares.
- Como entrada directa en un análisis de la exposición del prestamista hipotecario al riesgo de impagos.

Existen otras aplicaciones en la política monetaria y económica de los estados, ligadas al control de fenómenos del mercado inmobiliario que tienen efecto en la economía real:

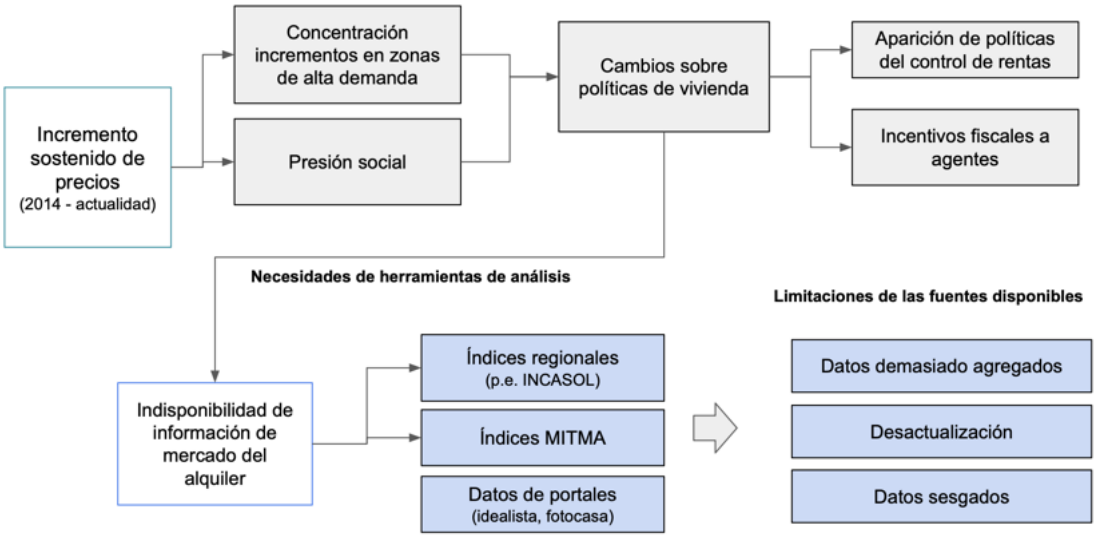
- Las burbujas de precios inmobiliarios (y los colapsos posteriores) se han relacionados con las crisis financieras y, por lo tanto, es importante identificar la formación de estas burbujas con precisión y de una manera

que sea comparable entre países, dado el impacto adverso de los procesos un crecimiento abrupto y excesivo (Case y Quigley, 2008) y anticipar las consecuencias de su re-equilibrio. El IPV es necesario para diagnosticar de forma temprana dichos problemas, ya que en muchas ocasiones, estos fenómenos no son explicables solamente con los factores fundamentales como los costes de construcción o precios de alquiler, sino que atiende a fenómenos psicológicos como apunta (Shiller, 2007a). Por otro lado, la política de tipos de interés de los bancos centrales pueden alimentar incrementos de precios en el largo plazo (Shiller, 2007b).

- Los índices de precios inmobiliarios son necesarios tanto para una correcta conducción de la política monetaria, como para establecer un control continuo de los mercados residenciales y los activos inmobiliarios comerciales.

Para que los índices incorporen la singularidad de la vivienda, Eurostat -(Eurostat, 2013) recomienda el uso de índices de precios hedónicos, que como su propio nombre indica, utilizan modelos de precios hedónicos para la construcción de los índices (Berndt y Rappaport, 2001).

Figura 2. Factores que determinan la necesidad de información en el mercado del alquiler español



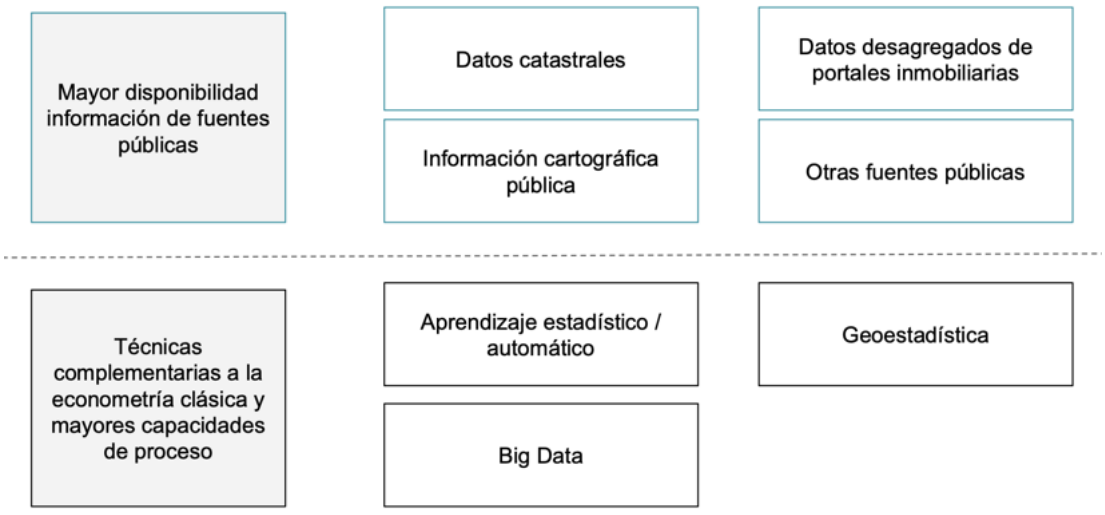
Fuente: elaboración propia.

Hasta hace unos años, la gestión de la política de la vivienda del alquiler a través de instrumentos analíticos se había considerado un objetivo secundario. Sin embargo, las condiciones de mercado de las últimas décadas ha revertido la cuestión, tal como ilustra la Figura 2. En el caso de España, el primer índice de alquiler se desarrolló en 2020, que a pesar de su nombre, en realidad no

es un índice de precios sino que es una serie histórica de medianas de precios. Esta situación cambió después de la gran crisis del 2008, a raíz de normativas como la europea que exige a las agencias estadísticas de los estados miembros la publicación a través de Eurostat un índice de precios armonizado de la vivienda trimestral (Eurostat, 2022), siguiendo una normativa común (que no significa que los datos de todos los países sean equivalentes). Estos instrumentos son clave para atender a uno de los activos más importantes de los países, tal y como puso de manifiesto el gobernador de la reserva central de Australia, Ian Macfarlane, en el año 2004 y que cita Hill (2013):

> *"... La vivienda es el mayor activo del país. Ciertamente, para el sector de los hogares representa alrededor del 60-70% de su riqueza total. Es una clase de activo extremadamente importante para la mayoría de las personas, sin embargo, la información que tenemos sobre los precios es desesperada en comparación con la información que tenemos sobre los precios de las acciones, los precios de los bonos y los tipos de cambio de divisas, e incluso obre los precios de las materias primas, los precios de exportación, importación y precios al consumidor. Realmente es probablemente el eslabón más débil de todos los datos de precios del país, así que creo que es algo en lo que me gustaría que se invirtieran recursos ..."*

Figura 3. Nuevas capacidades disponibles y datos alternativos para el desarrollo de índices del precio de la vivienda



Fuente: elaboración propia.

Por tanto, un índice de precios de la vivienda residencial del alquiler, preciso, actualizado y con un alto nivel de detalle, será instrumento de análisis esencial

para el control y el desarrollo de políticas las políticas del sector. Las cuales permitirán a la administración reaccionar de forma temprana a la formación de desequilibrios en el mercado. Idealmente, el índice de precios debe incorporar fuentes de datos relevantes de los distintos ángulos del mercado, elaboradas a través de nuevas tecnologías y metodología, como resume la Figura 3. Una de las ventajas, con respecto a los enfoques tradicionales, será su enorme nivel de detalle e inmediatez. Para ello, el uso de fuentes de datos alternativas, como los portales inmobiliarios, serán cruciales porque aportan una profundidad y nivel de actualización no disponible en los datos de organismos oficiales actuales.

Objetivos

El objetivo final de esta Tesis es el desarrollo de una metodología para la construcción de índices del precio de la vivienda en alquiler, con un alto nivel de desagregación funcional y temporal, que utilice como fuente principal los datos de oferta de portales inmobiliarios de Internet. La finalidad del trabajo puede desglosarse en tres metas principales, que a su vez se descomponen de una serie de objetivos secundarios representados en la Figura 4.

Figura 4. Objetivos de la Tesis Doctoral

Fuente: elaboración propia.

Las metas de la metodología atienden a la capacidad del índice de responder a las necesidades de análisis de sus posibles usuarios y asegurando en todo momento, en la medida de lo posible, que su información concuerde con la situación representada en los registros oficiales. Para ello se establecen cuatro

objetivos principales:

- Fiel reflejo del mercado del alquiler: los índices de precios del alquiler deben ser consistentes con los registros de fuentes oficiales, eliminando las distorsiones por sesgos de las fuentes de información de oferta utilizadas, principalmente la que procede de portales de Internet. Los cuales se componen por anuncios publicados por diversos agentes del mercado (particulares y profesionales), y que por tanto, no cuentan con criterios homogéneos en el establecimiento de precios. Tampoco hay un control exhaustivo de de anomalías y errores, y un mismo inmueble se puede estar publicado múltiples veces simultáneamente. Estos factores hacen que la muestra de oferta no represente fielmente al colectivo de inmuebles en alquiler, es decir, sobrerrepresente a una parte, infrarrepresente a otra y para algunas subpoblaciones no exista información. Para mitigar estos sesgos, se requiere conciliar los datos macroeconómicos sobre la composición y naturaleza del colectivo para ajustar los elevadores muestrales.
- Alto nivel de desagregación funcional: la evolución del precio se puede desglosar según varias características de los inmuebles de interés para los usuarios de la información. Dichas características incluyen las configuraciones constructivas, las clases de equipamiento con los que cuenta la vivienda, el tipo de edificación y la zona geográfica en la que se encuentra la vivienda. Este nivel de detalle permite el estudio de los precios para estratos relativamente pequeños, como por ejemplo: las viviendas de tipo unifamiliar, de cuatro habitaciones en municipios de entre 50.000 a 100.000 habitantes, con una antigüedad de menos de 25 años y en una zona urbana de lujo.
- Alto nivel de desagregación temporal: las series de índices públicos de vivienda se encuentran, generalmente, en frecuencias anuales o trimestrales, lo cual es insuficiente para una vigilancia continua del mercado. Un índice de mayor frecuencia, por ejemplo mensual, permite una mejor anticipación ante posibles cambios de tendencia o choques en el mercado de una forma. Además, las series de datos públicas suelen publicarse con un retraso de meses o años[14] que inhiben la inmediatez, por lo que se pretende generar series de datos mensuales cuyo último valor disponible sea el del mes inmediatamente anterior al actual.
- Uso de técnicas y fuentes de información alternativas: por una parte, el uso de datos sobre la evolución de los precios de oferta[15] de los portales

[14] El Índice de Precios de la Vivienda del INE se publica con un retardo de tres meses, y el Índice de Precios del Alquiler del Ministerio de Transportes dos años.
[15] Se define la oferta como todos los inmuebles disponibles para ser alquilados.

inmobiliarios aporta una perspectiva de los cambios del mercado inmobiliario en alquiler[16], lo que la convierte en una fuente candidata para el desarrollo de modelos que estimen el impacto en los precios del mercado. Por otra parte, el componente clave para la construcción de índice es el modelo de valoración de precios hedónicos de la vivienda[17], donde la ubicación geográfica será uno de los atributos más importantes, ya que la zona determina en buena manera los precios. Dada la gran cantidad de información de datos geográficos disponibles, se incorporan distintas fuentes públicas para la mejora de la precisión y expresividad de los modelos (Catastro, Instituto Geográfico Nacional, INE, Idealista). Por último, la metodología a desarrollar debe combinar las técnicas econométricas de construcción de índices habituales en la literatura y la industria, con técnicas de modelos de aprendizaje automático que permitan una mayor precisión y mejor capacidad de aprovechamiento de grandes volúmenes de información.

Junto con los objetivos principales de la investigación, además, se persiguen una serie de objetivos secundarios. En primer lugar, ligado al reflejo de la imagen del mercado, también se establecerá como prioritario que los índices de precios resultantes sean consistentes con las cifras oficiales del sector, de manera que se puedan utilizar los índices generados como una herramienta de validación de la estadística oficial y que permita inferir de forma adelantada estos valores. Un segundo objetivo secundario es que el modelado de la población, en estratos por características y zonas, se haga de la forma más detallada posible, hasta el punto de estimar el elevador muestrales de cada uno de los inmuebles de la muestra de oferta. La ponderación será crítica en la reducción de sesgos, y ayudará a establecer las relaciones funcionales (precios) entre las ambas poblaciones a lo largo del tiempo.

En cuanto al nivel de desagregación, los objetivos secundarios se refieren, por un lado, a los denominados funcionales: la capacidad del índice de disponer de un desglose por atributos constructivos o físicos de la vivienda, tipo de instalaciones de la misma o características socio-demográficas de la zona. Por otro lado, la desagregación debe poder realizarse criterios geográficos, o submercados de la vivienda[18], de forma que se logre el máximo desglose geográfico para el tamaño muestral mínimo para cada estrato. Además, el uso de modelos hedónicos en el proceso, permitirá conocer la contribución de las distintas características de la vivienda al precio, y ajustar dicha magnitud en el índice final en función de la

[16] El mercado del alquiler abarca la base de inmuebles en régimen de alquiler sobre los que los inquilinos pagan una renta.
[17] Un modelo hedónico supone que el valor de un bien se calcula como la suma de la contribución al precio de todos los elementos que lo componen
[18] Municipio y barrios.

calidad y características de cada vivienda.

Las técnicas de desagregación temporal permitirán desglosar las series de precios en valores mensuales, que el futuro podrían extenderse a frecuencias semanales o diarias. Por otra parte, se pretende que los modelos aprovechen la disponibilidad inmediata de los datos procedentes de los portales de Internet, para disponer de índices de precios actualizados de una forma casi en tiempo real.

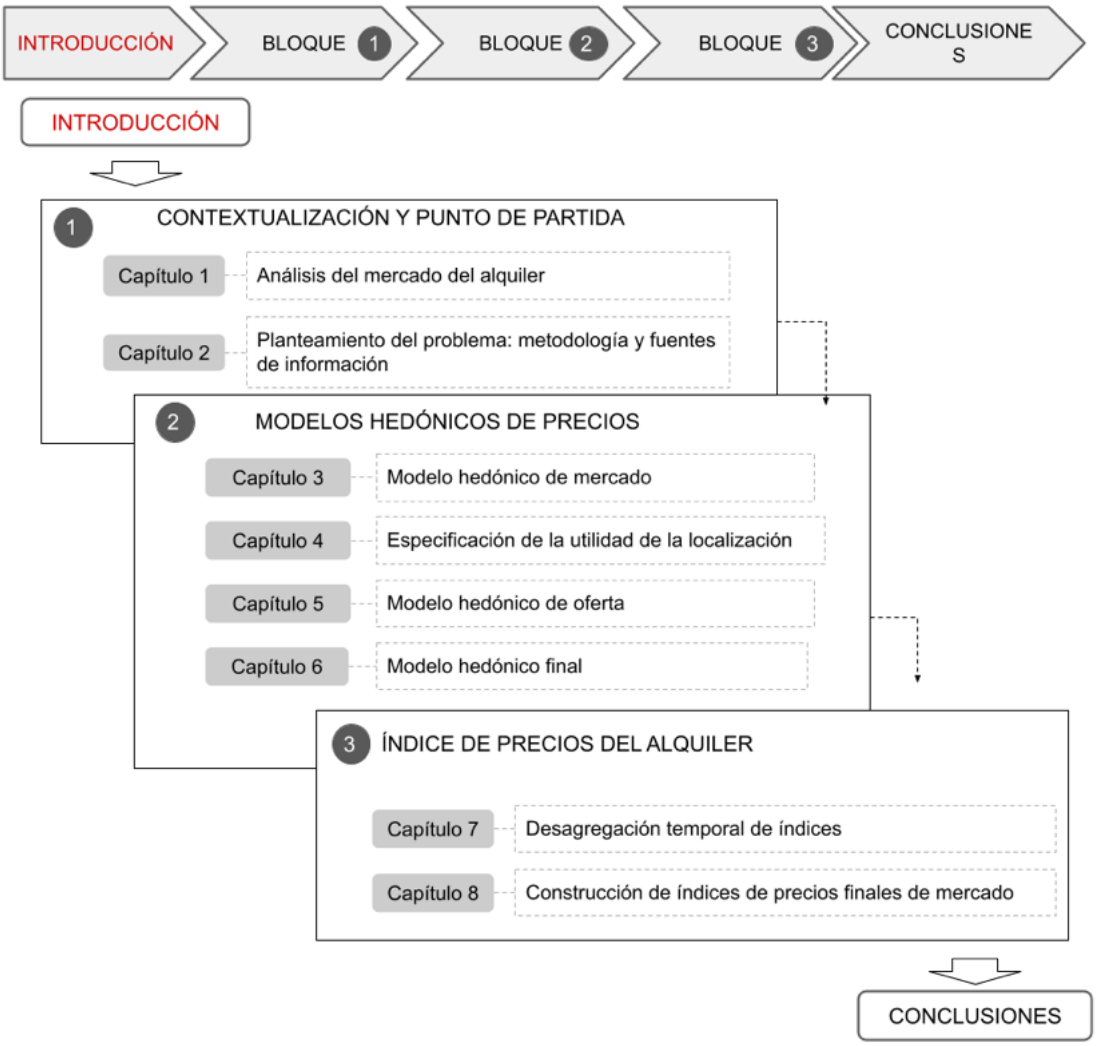
Otro objetivo secundario se centra en hacer uso extensivo de fuentes de datos abiertas y de Internet por varios de los motivos expuestos anteriormente: abundancia, alto grado de detalle de la información, actualización y fácil disponibilidad de la información. Es importante recalcar la importancia de las fuentes de información geográfica, constructiva y urbana del dato del Registro Central de Catastro (Direccion General del Catastro, 2022), y de la cartografía de uso abierto de Open Street Map.

Por último, la metodología tiene como meta secundaria la aplicación de técnicas de aprendizaje estadístico por tres motivos: 1) para reducir las limitaciones de los modelos de regresión lineal tradicional (multicolinealidad, heterocedasticidad, dependencia y heterogeneidad espacial); 2) aumentar el nivel de precisión de la estimación de los precios; y 3) para aprovechar el gran número de características de la vivienda que ofrece el repertorio de datos abiertos.

Estructura

Para cumplir los objetivos marcados, y considerando el marco y las circunstancias expuestas anteriormente, la presente investigación se articula en tres bloques. La estructura de la Tesis Doctoral sigue el orden de pasos de la metodología de investigación, los cuales se ilustran en la Figura 5. El Bloque I parte de un análisis del contexto inmobiliario histórico del mercado de la vivienda español, para posteriormente estudiar el marco teórico y las fuentes de datos que se utilizarán; pasando al Bloque II, que determina el colectivo del alquiler que representa la muestra sobre la que se construyen los modelos de precios hedónicos; finalmente, el Bloque III se centra en construir los índices de precios, partiendo de un primer índice con frecuencia anual que, a través de un proceso de desagregación temporal, da lugar a los índices de alquiler definitivos con frecuencia mensual.

Figura 5. Estructura de la Tesis Doctoral



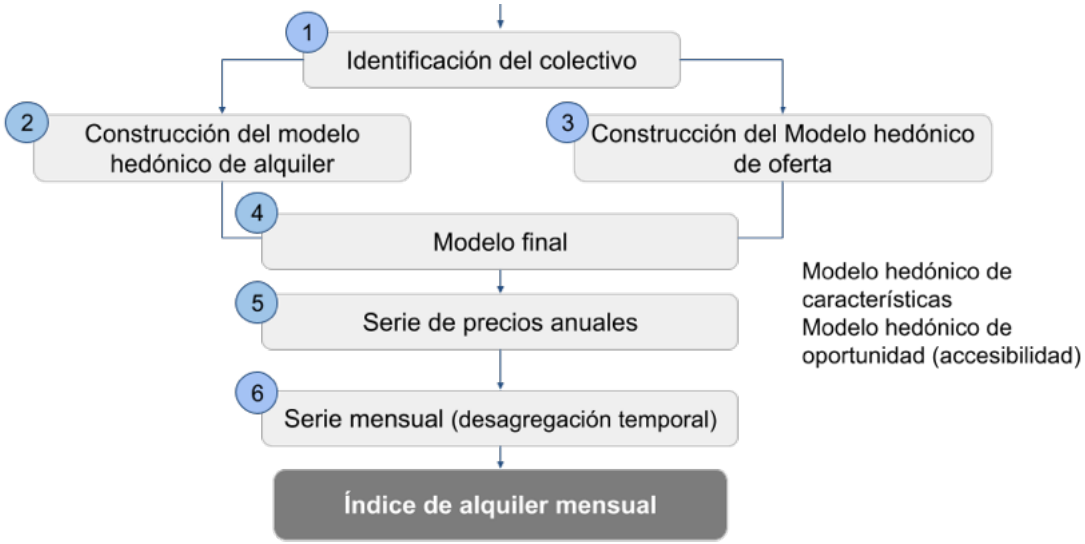
Fuente: elaboración propia.

La Tesis comienza con el el Bloque I, que describe el contexto del mercado de la vivienda del alquiler en España y plantea las cuestiones claves del problema, la metodología a emplear y las fuentes de información utilizadas. Consta de dos capítulos, el primero focalizado en analizar el contexto del mercado de la vivienda en España y en el ámbito europeo, que ahonda en el estudio pormenorizado de la situación actual e histórica del mercado del alquiler en nuestro país. Adicionalmente, se examinan las diferentes regulaciones del alquiler aplicadas tanto en España como en otros países europeos, atendiendo, principalmente, a las repercusiones que han tenido en los precios de las rentas.

El segundo capítulo del primer bloque describe el planteamiento metodológico y las fuentes de información utilizadas en la metodología. La cual se desglosa desde dos planos: el primero, el de los precios, centrado en explicar los factores que forman el precio de la vivienda; y el segundo, el temporal, enfocado en la

construcción de series de datos que reflejan la evolución temporal de los precios mediante números índices. Este segundo capítulo, además, describe cada una de las fuentes de información utilizadas para la investigación, con especial atención al tratamiento de errores y valores atípicos de la información procedente de portales inmobiliarios de Internet.

Figura 6. Descripción general metodología de índice de precios residenciales



Fuente: elaboración propia.

El Bloque II se enfoca en el desarrollo de los modelos hedónicos de precios de la vivienda en alquiler, y se compone de cuatro capítulos. El tercer capítulo, y primero del segundo bloque, se dedica al desarrollo del modelo del mercado de la vivienda en alquiler. Por tanto, se estima el precio que los individuos pagan por la renta de su alquiler, en contraposición con el precio de oferta o mercado que es el precio que pide el propietario en el portal de Internet de donde proceden los datos. La Figura 6 muestra el proceso completo que sigue la metodología, y que comienza con la construcción de varios modelos hedónicos que se aplican a la elaboración de los índices de precios anuales. El dato anual se descompone finalmente a frecuencia mensual a través de técnicas de desagregación temporal.

El cuarto capítulo se dedica al estudio de la especificación utilidad de la localización en los modelos hedónicos. La cual se sustenta en la premisa de que la ubicación es uno de los principales determinantes del precio, y que la prima de precio de una zona estará determinada por la utilidad percibida por el comprador en base el acceso a diversos servicios de la propia vivienda. Se desarrolla un método para la creación automática de variables asociadas de accesibilidad de servicios para cada una de las viviendas, basadas en la cercanía a puestos de trabajo, colegios, vías de transportes o servicios de restauración. Estas variables

de accesibilidad se utilizarán para especificar las particularidades de la ubicación de la vivienda, prescindiendo del uso de las coordenadas geográficas o nombres o códigos de barrios.

El quinto capítulo describe el desarrollo del modelo de alquiler de oferta, es decir, el precio por el que los propietarios o las agencias ponen al mercado sus inmuebles. Al contrario del modelo de mercado, este cuenta con una gran cantidad de variables, por este motivo se construyen dos modelos hedónicos complementarios. El primero de los modelos de oferta estima el precio a partir de las características físicas del inmueble, mientras que el segundo, utiliza la localización de la vivienda. Los dos modelos combinados, a través de un método denominado ensamblado, dan lugar a un estimador de precios que mitiga los inconvenientes de tratar con una gran cantidad de factores con estimaciones más precisas y robustas.

El Bloque II termina con el sexto capítulo, cuyo objeto es la construcción del modelo hedónico final. El cual parte del hedónico que estima los precios de mercado de forma muy agregada, y descompone desde un punto de vista funcional y geográfico dichos valores con la ayuda del hedónico de oferta. A este proceso se le aplica un ajuste adicional que corrige los sesgos de la estimación producto de una pobre especificación zonal en las estadísticas oficiales de mercado.

El Bloque III, se centra en la construcción de los índices de precios de mercado con frecuencia mensual y cuenta con dos capítulos. Se inicia con la creación de los índices anuales y su ajuste a frecuencia mensual, sobre los cuales se desarrollan los índices definitivos.

El objetivo del capítulo séptimo se dedica dedicado a la construcción de las series de precios mensuales, que a partir de los precios anuales, y mediante técnicas de desagregación, permiten la construcción de las series de precios de la vivienda mensuales. Dado que existe un gran número de técnicas de desagregación temporal, se desarrolla una metodología que persigue identificar automáticamente cuál es la mejor técnica de desagregación para cada serie. Cuestión que no es menor, dada la gran cantidad de series de precios a generar, y que hace inviable un proceso de selección manual de la técnica de desagregación más apropiada para cada serie. En el propio capítulo, se realiza un estudio del marco teórico de la desagregación de series temporales y las recomendaciones propuestas por Eurostat al respecto, las métricas de control de calidad de los procesos desagregación.

El octavo y último capítulo del tercer bloque se dedica a la construcción de los índices de precios finales del alquiler, tanto para precios de mercado como de

oferta. Se describen con detalle los criterios de estratificación (de tipo geográfico y funcional) y el cálculo de los números índices, que son del tipo Fisher encadenados. En este capítulo examina la calidad de las series generadas desde el punto de vista estructural, así como su consistencia en los ámbitos geográfico, temporal y de mercado. Para este último caso, se confrontan los resultados de la presente metodología con los valores del índice de precios de la vivienda experimental publicado por el INE. Finalmente, se evalúa la capacidad predictivas de las series de precios calculadas en futuras evoluciones del mercado inmobiliario.

La última parte de la Tesis Doctoral expone las conclusiones que se extraen de este trabajo. Con las mismas y la metodología propuesta se espera haber contribuido, aunque de manera incipiente, a ofrecer una herramienta que permita el análisis en detalle de las dinámicas de precios del alquiler, un método que, además, sea capaz de incorporar con mínimos sesgos fuentes de información abiertas en índices de precios de la vivienda. Por último, se espera que los métodos desarrollados ayuden al desarrollo de índices que garanticen la coherencia con las fuentes de datos oficiales con un alto grado de detalle, mediante de métodos innovadores de modelización de precios hedónicos.

Capítulo 1

Análisis del mercado de alquiler

> "No esperes a comprar un inmueble, compra un inmueble y espera."
>
> — Will Rogers, actor

1.1 Introducción

El mercado inmobiliario español alcanzó a inicios del 2019 un momento de máxima actividad, después de casi una década de depresión y recuperación[1]. En 2008, se inició un periodo de caídas de precios y descenso del número de compraventas que se extendió hasta 2014, y que por otra parte, tomó velocidades diversas en las diferentes zonas del país (Idealista, 2022). En ese mismo período, las rentas del alquiler entraron en una fase de mayor dinamismo y comenzaron la recuperación antes que el mercado de compraventa (Idealista, 2022). Todo ello ha llevado a una modificación gradual de la proporción entre los distintos regímenes de tenencia de la vivienda (compra y alquiler), tal y como como expone el Banco de España (2021).

La historia reciente del mercado inmobiliario en España explica el incremento del peso del alquiler. Entre 1997 y 2007, bajo el paraguas de una financiación masiva y barata, se produjo una explosión de la construcción y la compra intensa de vivienda. Apoyada en un intenso incremento de la población, principalmente por la inmigración, que alentó la demanda de la vivienda. Como consecuencia, se formaron importantes desequilibrios en la economía española, que la hicieron

[1] Excluimos de esta afirmación los efectos temporales de la crisis del Covid-19, por lo anómalo de la situación.

altamente dependiente de la financiación externa y la construcción se convirtió en su motor principal. Este periodo se caracterizó por la estrecha relación del sistema financiero con la administración regional y local. Existía un importante control de las Comunidades Autónomas sobre los órganos de gobierno de las Cajas de Ahorros, en ocasiones, con personal poco especializado. Las administraciones locales pusieron en mercado una importante cantidad de suelo público, que abrió un importante flujo de ingresos para los ayuntamientos y que aprovechó un sector inmobiliario ávido de construir.

Paralelamente, los precios de la vivienda crecían de forma continuada, alimentando de forma creciente una burbuja inmobiliaria que explotaría en 2008, a consecuencia de la exposición de la economía mundial a la crisis de las hipotecas subprime en 2007. La situación precedente llevó a una regularización, abrupta, violenta y repentina de los sectores financiero e inmobiliario, que dio lugar a una regeneración acelerada del sector bancario español con la virtual desaparición de las Cajas de Ahorro y con una parada, casi total, del crédito que afectó no solo a la vivienda, sino al resto de la economía. Esta situación de desapalancamiento, con un parque inmenso de vivienda nueva construida sin vender y la necesidad de un cambio de modelo productivo menos apoyado en la construcción, se prolongó entre los años 2008 y 2013.

A nivel social, tuvo importantes consecuencias en el número de desahucios en viviendas alquiladas y en propiedad, con un respaldo institucional mínimo, y bajo un marco regulatorio casi inexistente en políticas sociales de acceso a la vivienda. Quizá, en parte por el enorme crecimiento de la deuda pública, acrecentada por una crisis económico-financiero mundial que se magnificó en España por la explosión de la burbuja inmobiliaria. La situación se estabilizó finalmente a partir del 2014, con mejoras del acceso al crédito y crecimiento de la economía. El escenario macroeconómico y financiero post-crisis tuvo como consecuencia que parte de la población se viera forzada a optar por el alquiler como única opción de acceso a la vivienda.

El despegue del alquiler que comienza en 2008, coloca a España en vías de convergencia al modo de tenencia más habitual de los países de la Unión Europea. Este cambio no vino exento de efectos asociados, entre ellos, la presión al alza de los precios, que no fue un fenómeno exclusivamente español ya se produjeron incrementos en los precios en las zonas metropolitanas más importantes de las economías desarrolladas (Banco de España, 2020; Eurostat, 2022). Los motivos del incremento en la demanda se podrían explicar parcialmente por el cambio de hábitos de los ciudadanos, más proclives a alquilar que a comprar, y porque ciertos colectivos, con una mayor dificultad de acceso al crédito, tuvieron el

alquiler como la única alternativa habitacional. Adicionalmente, a finales de la década de los 2010 comenzaron a operar las grandes plataformas de alquiler vacacional[2] en los centros de ciudades turísticas, que redujeron la oferta del stock dedicado a rentas de largo plazo y, en consecuencia, condujeron a incrementos en los precios en estas localizaciones (Banco de España, 2020).

A pesar de la introducción de normativas en 2014, que desincentivan la compra de la vivienda en detrimento del alquiler, España no ha desarrollado medidas de intervención eficaces en el mercado para controlar el incremento continuado de los precios y garantizar el acceso a la vivienda.

Visto desde el ángulo micro-económico, la asimetría de información entre oferentes y demandantes de vivienda genera una escasez artificial. En parte por la falta de confianza en el arrendatario, y que se traduce en la aplicación de una prima uniforme sobre el alquiler independientemente de la solvencia del inquilino (Arruñada, 2022). Además, la inestabilidad jurídica causada por los continuos cambios normativos de las entidades nacionales, regionales y locales, sumada a una legislación que incentiva los alquileres pluri-anuales, restringen aún más la oferta, lo que inevitablemente conduce al incremento de los precios.

Las asimetrías de información son consecuencia directa de la ausencia de registros oficiales que recojan la evolución de las rentas a nivel nacional, regional y municipal. López (2007) apunta que la falta de información es producto de la dificultad de construir medidas comparables, dada la heterogeneidad de las características de la vivienda, y por la ausencia de un mercado central en el que se realizan las transacciones. Sólo en ciertos ámbitos locales o regionales, como el Gobierno Vasco, la Generalitat Valenciana, la Generalitat de Cataluña o el ayuntamiento de Barcelona existen bases de datos suficientemente detalladas. La escasez de información dificulta el conocimiento objetivo y exhaustivo de las dinámicas del mercado del alquiler, que aparte de ser información esencial para los agentes privados, es indispensable para articular una eficaz política pública de la vivienda.

Para mitigar la falta de información, el Ministerio de Transportes, Movilidad y Agenda Urbana, decidió publicar en 2020, una serie de datos de precios de alquiler (MITMA, 2020). Esta fuente, tiene una utilidad limitada por dos motivos, el primero es que no se trata de índice de precios sino que es un registro de medianas de precios, sin ajustes de composición o calidad; el segundo es que se publica con dos años de retraso.

Por otra parte, los portales inmobiliarios también publican información de la

[2] Entre ellas Airbnb, Booking o Homeway.

evolución del mercado, su ventaja es que están muy actualizados y cuentan con un gran desglose zonal. Sin embargo, estos datos representan el ángulo de la oferta no el mercado en sí mismo. Es decir, muestran la evolución de los precios de los inmuebles en comercialización, pero no las cantidades que se pagan como rentas del alquiler. El dato de los portales cuenta con otra limitación que es que la composición muestral de la oferta difiere de la del mercado en alquiler.

La preocupación por la evolución de los precios del alquiler a partir de 2018 y su impacto en colectivos de menores ingresos, han ampliado los desequilibrios del mercado a pesar de las distintas políticas de apoyo al inquilino, como apunta el Fondo Monetario Internacional (2021). Además, ha generado un fuerte debate social que ha producido la introducción de nuevas regulaciones sobre el alquiler. La intervención pública en el mercado del alquiler ha actuado en tres ejes: limitación de precios, promoción de vivienda social en alquiler y medidas indirectas que fomenten el alquiler privado.

Como es de esperar, una demanda de alquiler creciente y una oferta relativamente constante dan lugar a una tensión en los precios en las zonas con mayor demanda, como son los centros de las ciudades más importantes. Por tanto, los organismos nacionales, regionales y locales se ven empujados a intervenir para asegurar el acceso a la vivienda, principalmente a los segmentos de la población que por imposibilidad de acceso al crédito no se pueden permitir el acceso a una vivienda en propiedad. Gran parte de estas medidas, entre 2019 y 2021, han estado orientadas al control de los precios en determinadas zonas, como el caso de la ley de control los precios máximos de alquiler en Cataluña, aprobada en septiembre de 2020, y que fue el preludio de la ley de aplicación estatal aprobada en 2022 (sin embargo, en 2022, una sentencia del Tribunal Constitucional anuló los artículos principales de la ley catalana).

La situación inflacionista de los precios del alquiler no ha sido algo exclusivo del mercado español, sino que se ha producido en casi todas las economías avanzadas después de la crisis financiera del 2007. Tampoco lo han sido las políticas de limitación de precios, que se han ido implementado, con escaso éxito, en estas geografías. Por ejemplo, la aplicación en capitales importantes de la Unión Europea, como Roma, París o Berlín, no han logrando el objetivo deseado. Existe una larga literatura acerca del tema que se inicia desde el plano teórico (Bloomberg, 1947). La mayoría de estos estudios, como apuntan Diamond, McQuade y Quian (2019), constatan que la simple limitación máxima del precio (también denominado control de precios de segunda generación) (Arnott, 2003) ofrece resultados de reducción de precios nominales de las rentas en el corto plazo, pero ligados a un decremento en las unidades de oferta. Este impacto en

la oferta da lugar a un incremento en los precios reales a largo plazo, entre otros efectos. Recientemente, se empiezan a publicar estudios como los de Monrás y Montalvo (2022) o Arruñada (2022) que vienen a corroborar de forma preliminar los planteamientos conocidos de la literatura económica.

Para realizar un análisis más profundo, el capítulo se estructura en dos partes: un análisis del mercado del alquiler residencial europeo y español, éste último se analiza desde una perspectiva histórica los mercados de compraventa y alquiler, para finalmente estudiar su composición actual y los principales agentes que participan en él; la segunda parte analiza los mecanismos de gestión pública de control de los precios, y qué políticas se han aplicado recientemente en el ámbito estatal y europeo.

1.2 El mercado del alquiler

En España, la vivienda en propiedad ha sido la norma durante décadas, pero desde la crisis inmobiliaria en 2008, se ha visto un cambio hacia una mayor prevalencia del alquiler. Sin embargo, aún está lejos de la situación en los países centroeuropeos donde es el formato dominante. Al mismo tiempo, el alquiler vacacional a través de plataformas tecnológicas como Airbnb ha surgido en grandes ciudades turísticas, lo que ha producido una reducción en la oferta de alquiler residencial de larga duración y ha provocado, posiblemente, un aumento en los precios.

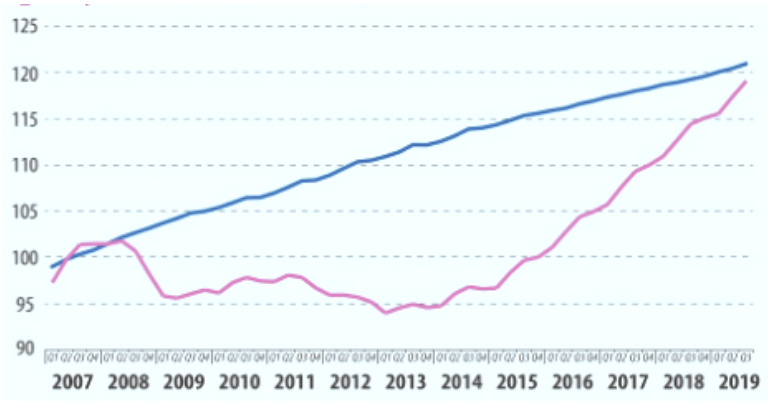
En la Unión Europea, el mercado de alquiler residencial ha experimentado un crecimiento en los últimos años, con una demanda sostenida por los ciudadanos que buscan una vivienda asequible y flexible. Al igual que en España, las ciudades europeas más grandes y turísticas, como París, Londres, Berlín y Amsterdam, han experimentado un mayor aumento en el precio de los alquileres, debido a la demanda por parte de los turistas y los jóvenes profesionales.

1.2.1 El mercado del alquiler en la Unión Europea

Los precios de la vivienda y los alquileres en la Unión Europea (a continuación UE) han seguido caminos muy diferentes desde el inicio de la última crisis financiera. Si bien los alquileres aumentaron de manera constante desde el inicio de la crisis, hasta el tercer trimestre de 2019, los precios de compraventa no lo han hecho así y han tenido periodos de importantes fluctuaciones, como observa la Figura 1.1. Según datos de Eurostat (2021, 2022), después de una fuerte caída inicial en 2008, a raíz de la crisis financiera, los precios de la vivienda se mantuvieron relativamente estables entre 2009 y 2014, para volver a aumentar

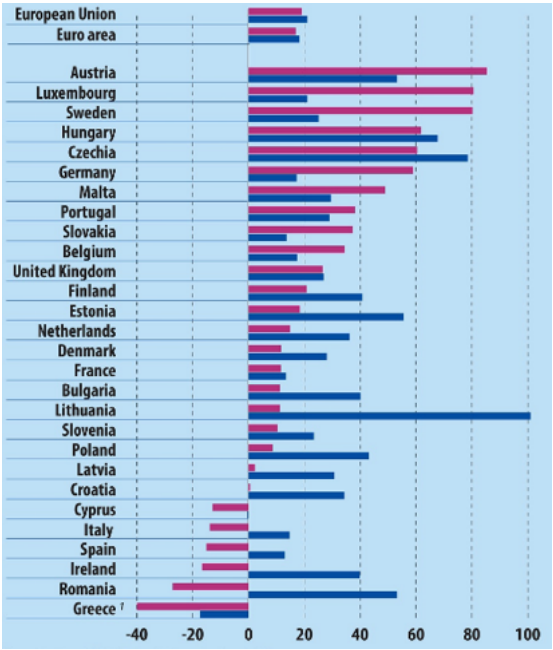
de forma continuada desde principios de 2015. En el periodo de 2015 a 2019, se aprecia que los precios de compraventa de viviendas aumentaron a un ritmo mucho más rápido que los alquileres. Para el periodo completo desde 2007 a 2019, los precios del alquiler residencial aumentaron un 21% en la UE, mientras que los precios de compraventa aumentaron sólo un 19,1 %.

Figura 1.1. Evolución de los precios de la vivienda y alquileres en la UE periodo 2007-2019



Fuente: Eurostat (2021).

Figura 1.2. Variación de los precios de la vivienda y alquileres en la UE por país



Fuente: Eurostat (2021).

Sin embargo, en este periodo, la evolución no fue homogénea en todos los estados miembros de la UE como se observa en la Figura 1.2. A excepción de Grecia, el resto de países muestra crecimientos de los precios alquiler, al igual que en

compraventa, con las excepciones de España, Italia, Rumanía, Chipre y de nuevo Grecia. Existen diferencias notables en la zona. Hay países dónde aumentan los precios de forma elevada como Lituania, con un crecimiento superior al 100 %. Mientras que en los países del sur, de forma general, los precios de compra descendieron cuando que los de alquiler aumentaron. Los países del norte y centro de Europa por contra, muestran incrementos en ambos mercados. Cabe destacar de nuevo el caso de la República Checa, con un aumento del 80 % del precio del alquiler y un 60 % del precio de venta, o el del incremento del 60% en el precio de venta de Alemania.

En términos de tenencia, en 2018 una cuarta parte (24,9%) de la población de la UE-27 vivía en una vivienda en propiedad, con una hipoteca o préstamo, mientras que más de dos quintas partes (45,1%) de la población vivía en una vivienda ocupada por el propietario, sin un préstamo o hipoteca. Alrededor de una quinta parte (20,8%) eran inquilinos con un precio de mercado de alquiler, y aproximadamente una décima parte (9,3%) eran inquilinos en alojamiento gratuito o de alquiler reducido.

Comparado con los países de la Europa occidental, España tiene una de las tasas más altas de propietarios sin hipoteca pendiente, después de Suecia, y la segunda más alta tasa de propietarios con o sin hipoteca (el 78,9% en 2012, el 77,7% en 2013 y el 76,3% en 2018). A lo largo de los últimos diez años, se ha producido un ascenso progresivo del porcentaje de ciudadanos que viven en régimen de alquiler, aún así se encuentra muy lejos de la media europea. Sólo los estados del este tienen tasas de tenencia en propiedad más altas que las de España. Según (Eurostat, 2022) en 2018 más de la mitad de la población de cada Estado miembro de la UE vivía en viviendas en propiedad. También en este aspecto se aprecian grandes diferencias entre países, véase el caso de Alemania con un 51,4% de propietarios o Rumanía con una tasa del 96,4%.

Los países bajos y nórdicos, en general, poseen mayoritariamente regímenes de tenencia de alquiler, y poco más de la mitad de las personas vivían en viviendas en propiedad o con una préstamo hipotecario. En cifras, en los Países Bajos el alquiler representa un 60,5%, Suecia un 51,7%, Islandia un 63,9% (datos de 2016) y Noruega con un 60,1%.

La proporción de personas que vivían en viviendas alquiladas, con un alquiler a precio de mercado en 2018[3], fue inferior al 10,0 % en 11 de los Estados miembros de la UE. Aunque se aprecian importantes diferencias en función del país, este formato representaba alrededor de dos quintas partes de la población en Alemania (40,8%); un 39,4 % en Dinamarca; más de una tercera parte en Suecia (35 %); un

[3]Fuente: Housing in Europe (Eurostat 2021).

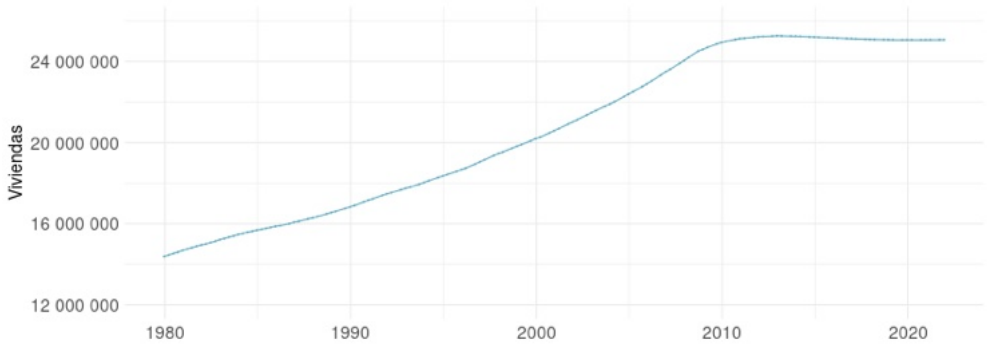
30,2 % en Países Bajos; un 29,7 % en Austria; una quinta parte en Luxemburgo (23,4 %); un 21,3 % en Grecia; un 19,4% en Bélgica, y Suiza representa el caso más extremo, donde superó la mitad (51,1 %). En este mismo año, en todos los estados de la Unión Europea, menos del 20% de las personas vivían en hogares con un alquiler a precios reducidos.

1.2.2 Evolución histórica mercado inmobiliario en España

El stock de viviendas en España aumentó alrededor de un 20% durante en la primera década del siglo XXI, según datos de INE (2011). Dicho stock de viviendas pasó de 20.946.554 viviendas en 2001 a 25.208.623 viviendas en 2011, tal y como se muestra la Figura 1.3. Los periodos de auge de la construcción y la demanda de vivienda más recientes fueron, 1970-1974, 1985-1990 y 1997-2007, siendo el último de ellos el más largo y abrupto en su finalización.

Entre 1997 y 2007, los precios de compraventa de vivienda aumentaron a un ritmo más rápido que la inflación: los precios reales de la vivienda se duplicaron en términos reales, evidenciando incrementos anuales de alrededor del 15 %. En 2007, justo antes del cambio de ciclo económico, el precio medio de la vivienda equivalía a nueve veces el salario bruto anual medio, un incremento sustancial si se tiene en cuenta que esta relación era de cuatro veces en 1997 (Rodriguez-López, 2009)

Figura 1.3. Parque de viviendas estimado



Fuente: Banco de España (2022).

Entre el año 2008 y el 2013 se produjo un colapso del sector inmobiliario en España, con una parada casi completa de la construcción y un cierre del acceso al crédito inmobiliario y promotor-constructor. Esto generó dificultades de acceso a la vivienda en propiedad en los colectivos de jóvenes y clases menos favorecidas, que alimentó la demanda, casi obligatoria, de la vivienda en alquiler. El incremento de la demanda en alquiler no fue compensado con el exceso de

oferta de obra nueva no vendida al final de la crisis, al menos en los grandes zonas metropolitanas, donde los precios subieron en el periodo 2014-2019 de forma notable.

A partir de 2013 se introdujeron cambios en la política fiscal de la vivienda residencial para los particulares, eliminándose las desgravaciones de los créditos de vivienda en el impuesto de la renta de las personas físicas, y la creación de deducciones por el alquiler de la vivienda. Estas medidas intentaban limitar el crecimiento del saldo vivo de crédito de la vivienda, así como reducir la ratio de tenencia de primera vivienda en propiedad, para limitar el peso de la deuda asociada a este bien. Sin embargo, estas medidas intensificaron aún más la demanda de la vivienda en alquiler, que unido a una oferta de alquiler pública y privada estrecha e inadecuada (Montalvo, 2011), con un casi ausente papel regulador del sector público para garantizar el acceso de la vivienda, derivó en la introducción de una serie de normativas orientadas a controlar la oferta. Entre ellas se encontraban la introducción de beneficios fiscales para el arrendador y la limitación de los precios de oferta.

Los mercado de alquiler y compraventa no se comportaron de igual manera durante el período de expansión. Mientras que en la compraventa se produjo una mayor especulación sobre los rendimientos futuros, en el alquiler, los rendimientos estrechos e inadecuados fueron definitivamente desatendidos desde un punto de vista regulatorio (porque la administración recibía grandes ingresos por la promoción de nueva vivienda, venta de suelo público y tasas asociadas a la compraventa). Se puede afirmar que este es uno de los resultados más negativos de un período de auge explosivo en el sector inmobiliario, y tuvo unas claras consecuencias negativas tras el estallido de la burbuja de los precios de la vivienda en 2007, como describe Rodríguez (2017) .

1.2.2.1 El auge (1997-2007)

La economía española tuvo una época de gran crecimiento en los años sesenta del siglo XX, con un aumento importante del precio de la vivienda y con ajustes a la baja en décadas posteriores. Hay que destacar los desarrollos inmobiliarios de los años 1971-1974 y los realizados entre 1985 y 1990. Las fases de fuerte expansión de la demanda interna, asociadas al crecimiento de la construcción de tipo residencial, han estado acompañadas de elevados déficits externos o de balanzas de pagos por cuenta corriente. Esto se debe, entre otros motivos, a que el intenso crecimiento de la construcción residencial desvió los recursos productivos hacia este sector en detrimento a sectores productivos cuyo fin es la producción de bienes y servicios. En términos financieros, un déficit exterior elevado implicó

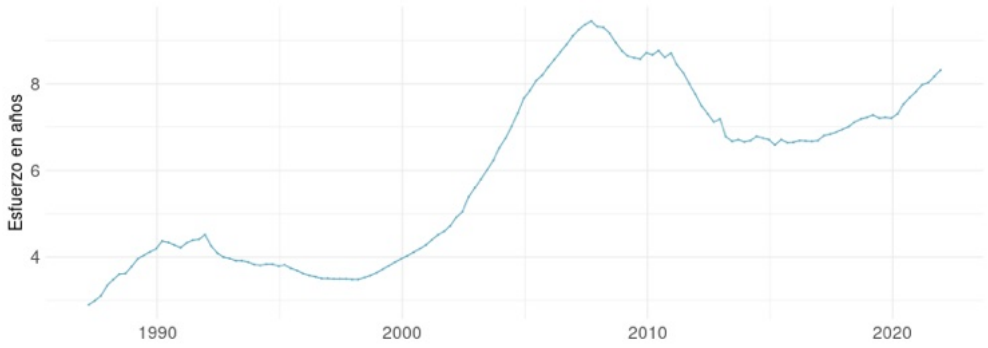
un mayor nivel de endeudamiento con el exterior, tanto privado como público, lo que provocó una canibalización de los recursos productivos por parte del sector inmobiliario, y que lastró negativamente la capacidad de competir exteriormente a la economía española.

El déficit exterior y la disponibilidad de crédito alimentó la burbuja inmobiliaria del periodo 1997-2007. Al incrementarse de forma notable el endeudamiento externo, lo que dio lugar al aumento de los tipos de interés pagados a pagar por la deuda.

El incremento de la demanda presionó los precios ascendentemente, que junto a un acceso casi ilimitado del crédito y a unas fuertes expectativas de beneficios, impulsaron de forma importante el crédito promotor, por tanto la deuda del triángulo: promotor, constructor y comprador adquirieron un peso muy relevante en la fase de expansión de crédito en esta época. Es importante destacar que los mayores impulsos del mercado inmobiliarios suelen derivarse de una expansión del crédito (Carbó Valverde y Rodriguez Fernandez, 2010).

En el periodo 1997-2007 el número de operaciones de compraventas fue elevado, alcanzando el máximo de la serie histórica en España en 2005, según datos del Banco de España (2021). Por otra parte, en 2006, se registró el valor más alto de vivienda iniciadas (865.000), según datos del MITMA (2022a), dando lugar a que en 2007 el nivel de construcción residencial alcanzara el 12,4% del PIB, según la Contabilidad Nacional (INE, 2022a). Durante esta década, los precios de compraventa crecieron anualmente 11,4 % de media en términos nominales (datos ministerio de MITMA (2022a)). En lo que se refiere a accesibilidad de la vivienda, la situación empeoró notablemente, con un promedio de 4,5 salarios brutos anuales para la compra de una vivienda de 90 metros cuadrados construidos en 1997, para alcanzar una media de 9,2 años en 2007, como se observa en la Figura 1.4.

Figura 1.4. Esfuerzo de compra de la vivienda en años, calculado como precio de la vivienda libre/renta bruta por hogar



Fuente: Banco de España (2022).

El contexto socio-político fue propicio a esta dinámica, el papel de los diferentes agentes alentó de forma crucial la formación de la burbuja. Los ayuntamientos recibían más ingresos por impuestos por la venta de suelo público, los promotores recibían crecientes e importantes beneficios de su inversión, y los compradores conseguían importantes plusvalías en sus inversiones. De hecho, en esta época la opinión general se sustentaba en creencias como “alquilar es tirar el dinero” o “los pisos no pueden bajar de precio”, bien resumido por el anuncio de radio de la época “No tire el dinero, compre” (Montalvo, 2011). Durante el proceso se publicaron distintos estudios advirtiendo de este fenómeno, en particular en 2003 el Fondo Monetario Internacional (Internacional, 2003) advirtió que el impacto de las burbujas inmobiliarias fue muy superior al de las equivalentes en el mercado financiero. En dicha época, la opinión de los expertos y los ciudadanos se dividía en aquellos que admitían la existencia de una burbuja y los que la negaban.

Una política monetaria generosa, por parte de los bancos centrales, trajo consigo tipos bajos de intervención, que se trasladó a los mercados mayoristas y posteriormente a los mercados minoristas, principalmente en los créditos inmobiliarios. Una liquidez suficiente procedente generalmente del exterior, una gran número de entidades financieras compitiendo por los clientes con unos criterios de riesgo generosos, dieron lugar a la concesión de un gran número de créditos.

Tanto la Reserva Federal de Estados Unidos como el Banco central Europeo en Europa, miraron para otro lado mientras crecía el crédito, muy particularmente en los mercados periféricos de la UE (España, Portugal, Italia, Grecia e Irlanda), bajo la tesis de que sus déficits exteriores solo tenían impacto en sus propias economías y no contaminarían al unión monetaria, premisa que se probó falsa dado que el déficit externo es siempre sinónimo de un aumento del endeudamiento externo.

Tampoco las entidades supervisoras de la actividad bancaria controlaron la situación, siendo los bancos centrales los principales responsables con una política laxa en la concesión de crédito; y tolerando déficits prolongados en la balanza de pago. Y no aplicando los instrumentos que tenían a su disposición: provisiones, coeficientes de caja o relaciones préstamo/valor.

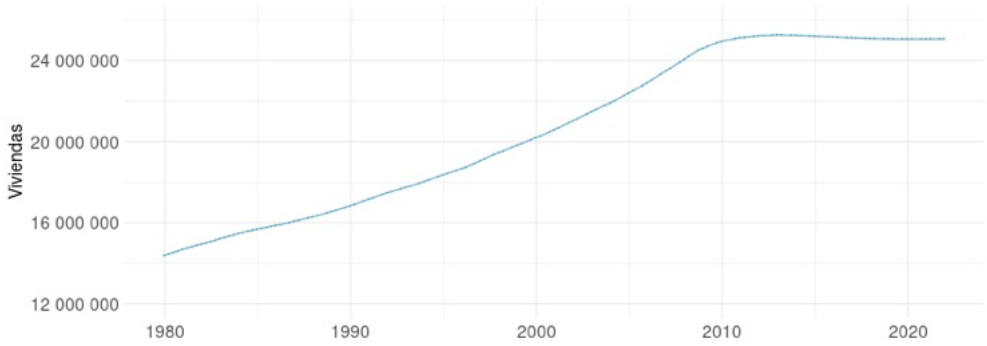
Los indicadores relativos a las magnitudes monetarias publicados por el Banco de España mostraron un enorme ciclo expansivo, de ahí que pasara de un tipo medio de interés de un 16,7 % para la compra de la vivienda, en 1990, a un 3,28 % de media en 2005. Esta coyuntura disparó la concesión de préstamos hipotecarios de forma importante, de 479.300 en 1997 a 1.342.200 en 2006, Banco de España (2021). Como resultado, el saldo vivo de crédito inmobiliario también creció sensiblemente a un ritmo anual del 22,5 % en el periodo de estudio (Banco

de España, 2022a).

No solo se produjo un acusado crecimiento del crédito a particulares, en un 20 % anualmente, sino que el crédito promotor creció a ritmos del 33 % al año (Banco de España, 2022a). El mayor crecimiento del crédito promotor con respecto al comprador anticipó dificultades de la venta de las futuras unidades construidas; y aunque en el conjunto de la UE se produjo una expansión crediticia inmobiliaria, España resultó ser una anomalía que crecía en más de 8 puntos porcentuales por encima de la Eurozona, según datos estadísticos del Banco Central Europeo (2022).

Como se aprecia en la la Figura 1.5, la exposición general de la economía al crédito inmobiliario también se disparó en este periodo, pasando la ratio entre el saldo vivo y el PIB de un 27,9 % en 1997 a un 100 % en 2007. Como consecuencia, el endeudamiento familiar ascendió desde el 35,4 % del PIB en 1997 al 83,6 % en 2007, (Banco de España, 2022a). Otro fenómeno de este periodo fue el incremento de los ingresos de los ayuntamientos, procedentes de los impuestos asociados a la construcción residencial. Esto detrajo recursos de las entidades locales a estas actividades en detrimento de otras de interés general de los ciudadanos. Los crecientes ingresos de las entidades locales, sumado a la numerosa presencia de los representantes de las administraciones competentes en el gobierno del suelo (urbanismo), ayuntamientos, autonomías en los órganos de gobiernos de las cajas de ahorro, impulsaron con fuerza la concesión de crédito promotor.

Figura 1.5. Crédito hipotecario. Total. Saldo vivo en porcentaje del PIB



Fuente: Banco de España (2022).

Un factor añadido fue el impacto demográfico de la inmigración a finales de los 90 e inicio de los 2000, pasando España de ser un país sin apenas inmigrantes, con un 1,6 % en 1998, a un 10 % en 2007, todo esto produjo un incremento significativo de la demanda.

Desde el plano regulatorio, la Ley 6/1998 señala la necesidad de incrementar de

forma importante el suelo urbanizable para reducir los precios del suelo y de la vivienda. Esta normativa estatal fue con gran probabilidad terreno abonado para la posterior espiral de especulación y burbuja de los mercados locales de la vivienda. El crecimiento de las vivienda en la última mitad del siglo XX fue enorme, desde los 6,7 millones de viviendas en 1950 se pasó a 25,2 a 2011. Como aspecto endémico, el mercado español guarda una elevada relación entre número de viviendas por unidad familiar, con 1,4 en España y 1,2 en el resto de Europa (Banco Central Europeo, 2022; Eurostat, 2021), y una reducida cantidad de viviendas destinadas al alquiler (en 13,5 % en 2011, frente un 36 % de media en la Eurozona) (Eurostat, 2021).

La economía española retornó al déficit exterior corriente, impulsado por la fuerte demanda interna de la economía, en la cual la construcción fue el principal dinamizador. Entre 2000 y 2011, la balanza de pagos mostraba déficits por cuenta corriente importantes que fueron especialmente agudos entre 2005 y 2008, en este último año se alcanzó un máximo del 10 % sobre el PIB. Esto produjo un aumento del endeudamiento frente al resto del mundo, que no sirvió para financiar el sistema productivo de la economía española sino para acumular una gran cantidad de viviendas sin vender.

1.2.2.2 El hundimiento (2008 - 2013)

El inicio de la gran crisis se desencadenó en Estados Unidos en 2007, debido a la titulización (enajenación de créditos a través de títulos) y venta masiva de créditos hipotecarios *subprime* (más de 500.000 millones de dólares) que se demostraron en su mayoría fallidos. Debido a la exposición de los mercados financieros mundiales a estos productos, la crisis tuvo impacto global. En España impactó en 2008 con la restricción del crédito, debido al cierre de los mercados de capitales, siendo la quiebra de Lehman Brothers el punto álgido de la crisis.

El contexto macroeconómico de la economía española cambió de signo a partir del 2008, registrando incluso decrecimientos del PIB desde el segundo trimestre del 2008, e importantes descensos del empleo que se prolongaron hasta 2013. La restricción del crédito también se tradujo a una fuerte restricción del crédito de las entidades bancarias españolas. Esta crisis financiera se sustanció en una crisis económica, que posteriormente se convirtió en una crisis de deuda y del euro donde los gobiernos intervinieron para mantener estabilizado el nivel de actividad y para evitar la quiebra de los bancos. La intervención se articuló mediante una fuerte emisión de deuda pública, que cada país realizó de forma aislada, lo que generó una fuerte crisis del euro, con su máximo exponente en las situaciones de tensión en Grecia.

Los bancos españoles habían recurrido al apalancamiento para mejorar sus cuotas de mercado y los resultados, los plazos cortos de los pasivos con largos plazos de sus activos (préstamos) generaron situaciones de desequilibrios en sus balances, aumentando su riesgo de crédito y liquidez. Por consiguiente, se inició un proceso abrupto y rápido proceso de desapalancamiento, las entidades frenaron casi totalmente la concesión de créditos y comenzaron a adelantar la recuperación de los créditos vivos.

En el verano de 2007, se inició un aterrizaje intenso de un mercado inmobiliario acelerado, la restricción del crédito redujo de forma importante el número de operaciones compraventa, en mayor medida en las de nueva construcción que en las de segunda mano. Esta fase de ajuste fue más acusada (INE, 2023b), entre 2011 y 2013, en paralelo al proceso de ajuste del sector financiero, reduciéndose un 37 % en número de operaciones entre el 2014 y el 2007, y en un 30,4 % en precio. Según la tasadora Tinsa (2023), el descenso de precios fue especialmente intenso en la Costa Mediterránea con casi un 50 %, mientras que el más moderado fue en Balearias y Canarias en un 30 %.

La nueva situación de falta de crédito unida a un exceso de oferta de obra nueva sin vender, dio lugar a la reducción del 90 % de la actividad de la construcción entre 2006 y 2012 (MITMA, 2022a), y redujo de forma notable el peso de la construcción residencial con respecto al PIB, según los datos de la Contabilidad Nacional (INE, 2022a). Todo ello llevó a un ajuste mucho abrupto que en crisis anteriores.

El frenazo repentino y masivo del mercado inmobiliario y el sector de la construcción, sumado al peso del mismo en la economía nacional, impactó en su crecimiento, pasando de contribuir en un punto porcentual en 2006 a detraer un punto entre 2008 y 2014.

En el segundo trimestre del 2013, la economía española alcanzó su mínimo en la recesión que comenzó en 2008. Y partir del 2014 comenzó su senda de crecimiento, aún así la erosión en el producto interior fue enorme, siendo el PIB en el primer trimestre de 2016 un 3 % inferior al alcanzado al segundo trimestre del 2008. Según el (2022a) el impacto de la recesión en la actividad fue mucho menor que su impacto en el empleo (INE, 2023c), que trajo consigo una disminución del 16 % del número de ocupados y unos 3,3 millones de empleos perdidos. La disminución del empleo en el sector de la construcción representó el 48 % de los empleos perdidos. Esta prolongada recesión significó que si en 2007 había 42,5 empleos por cada cien habitantes, en 2013 esta cifra era de 35,5, que representaba en niveles cercanos a 1997.

El sector primario y secundario resultaron muy afectados por la caída del sector

inmobiliario, y continuaron perdiendo peso a partir del 2007. De la misma forma, el sector público se vio afectado por el fuerte incremento de la deuda pública (de un 36,2 % del PIB en 2007 hasta un 100 % en 2016) (Banco de España, 2022a). Finalmente, el endeudamiento exterior dejó de crecer en este periodo y, por consiguiente, el déficit exterior se redujo entre 2007 y 2012.

Después del 2007 se produjo un retroceso relativo del crédito en el sector privado con respecto al PIB, y de manera mucho más acentuada el crédito hipotecario, pasando el saldo vivo con respecto al PIB de un 94,3 % en 2007, a un 63,1 % en 2015, (Banco de España, 2022a).

La recesión de la economía española, entre los años 2008 y 2010, estuvo estrechamente relacionada con una caída del mercado de la vivienda y la parada del sector de la construcción. Este retroceso se prolongó debido a las políticas de austeridad establecidas en la Eurozona para reducir el impacto negativo del elevado nivel de endeudamiento. El retraso del ajuste de los balances en el sistema financiero español hasta el 2015 retrasó la recuperación del mercado inmobiliario, lo que impactó enormemente al ecosistema bancario español. Como consecuencia se inició un proceso de concentración bancario, que llevó a una práctica desaparición de las cajas de ahorros.

En esta situación de recesión que produjo un aumento del desempleo y dificultad de acceso al crédito, la morosidad creció de forma extraordinaria, muy particularmente en lo que se refiere a crédito promotor y constructor, con unos niveles del 30,6 % frente al 10,3 % del resto de sectores de actividad. El impacto más visible, desde un punto de vista social, fue el número de desahucios de propietarios insolventes por ejecución de un préstamo hipotecario. El número de procedimientos de este tipo pasó de 17.605 en 2006 a 91.622 en 2012 según estadísticas del Consejo General del Poder Judicial (2012).

El fuerte incremento del paro dio lugar a un enorme crecimiento del esfuerzo de las familias para acceder al pago de una vivienda (Rodriguez López, 2017). En 2008, la razón de los costes de la vivienda con respecto a los ingresos anuales de una familia media española alcanzaron el 52 % (Rodriguez-López, 2009). Este desequilibrio provocó importantes dificultades en el acceso a la vivienda en propiedad de ciertos colectivos: jóvenes o familias con menores rentas más bajas, y que además se agudizó en las grandes ciudades. Como resultado, estos grupos tuvieron que recurrir al mercado de alquiler como única alternativa habitacional. Esta intensa demanda, en un mercado con una oferta estrecha, y con una fiscalidad que desincentivaba la compra, produjo un fuerte incremento de los precios de las rentas para el periodo entre el 2007 y el 2018. El detalle de la interrelación entre todos estos factores se explica en la Figura 1.6.

Figura 1.6. Factores explicativos del repunte del precio de la vivienda entre 2014 y 2019

Fuente: Banco de España (2020).

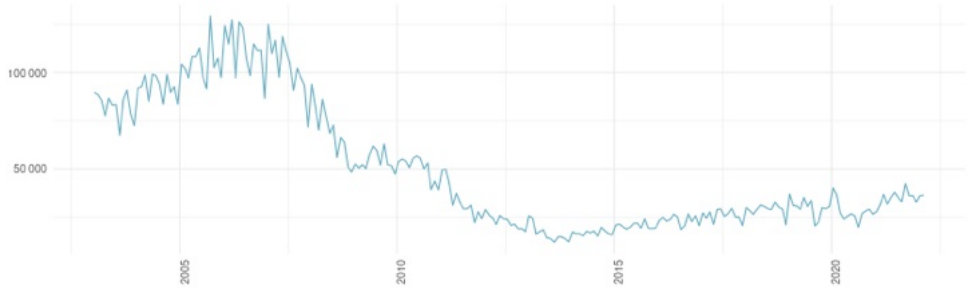
Los factores anteriores dieron lugar al replanteamiento regulatorio en años sucesivos, aunque esta cuestión no fue exclusiva del ámbito español sino en el europeo, como veremos en la sección dedicada a una revisión de las propuestas de políticas públicas. Además, el fenómeno de tensionamiento de precios[4] del alquiler de vivienda no es general, sino que se afecta principalmente en zonas metropolitanas, que históricamente han sido mercados de alta demanda y con una oferta limitada. La tensión de los precios puede relacionarse también con el comportamiento de la oferta en determinadas zonas, en particular en aquellas zonas con alta demanda y una oferta restringida.

[4]La tensión de precios de la vivienda se produce cuando el coste de la vivienda principal supera un porcentaje del total de los ingresos familiares.

1.2.2.3 La expansión (2014 - 2020)

En 2014 comenzó una nueva fase expansiva del mercado de la vivienda de uso residencial, tanto en compraventa como en alquiler (Román *et al.*, 2020), como queda patente en los datos estadísticos de la oferta y de las transacciones registradas (Idealista, 2022; INE, 2022b). Este fenómeno coincidió en el tiempo con una reducción continuada del tipo de interés medio de las hipotecas constituidas, pasando de tipos superiores al 4 % en 2014 a un 2,5 en 2020, (Banco de España, 2022a), como se muestra en la Figura 1.7.

Figura 1.7. Tipo de interés medio al inicio de las hipotecas constituidas (2014-2022)



Fuente: INE (2022).

El mercado de los propietarios de vivienda residencial ha sido tradicionalmente dominado por inversores minoristas, tendencia que se ha ido reduciendo en los últimos años con la entrada de inversores institucionales. Aunque actualmente, el mercado de la vivienda residencial en España sea principalmente en propiedad, esto no ha sido así en el pasado. Quizá el evento que desencadenó un cambio de tendencia, fue la aprobación de la Ley de Arrendamientos Urbanos en 1964. Esta ley introdujo prórrogas obligatorias de los contratos de arrendamiento que favorecieron a los inquilinos, les dio a los descendientes de los inquilinos el derecho a asumir sus derechos y obligaciones y también prohibió a los propietarios ajustar sus alquileres, de acuerdo con los precios del mercado (Rull, 2018). La ley impidió efectivamente que España crease el tipo de mercado de alquiler profesional, que se encuentra en otros países de Europa. En los últimos años, a consecuencia de una legislación más equilibrada, junto con un clima económico y un mercado laboral que han obstaculizado el acceso a la propiedad de la vivienda, han elevado el número de inquilinos en España.

Según datos del INE, el porcentaje de inquilinos en el mercado de la vivienda, subió de alrededor del 20 % en 2008 al 24 % en 2018. Sin embargo, a pesar de este repunte, España sigue estando lejos de sus homólogos europeos.

El perfil del inquilino actual de esta etapa no tenía un rango de edad concreto,

aunque se produjo un aumento de esta opción entre los jóvenes, según de la encuesta de condiciones de vida de 2019 (INE, 2019) el 52,4 % de personas entre 16 y 29 años alquilaba en 2018, frente al 36,5 % en 2008. Esto se debía en buena manera al hecho de que los jóvenes que abandonaban el hogar familiar, se veían obligados a recurrir al mercado de alquiler porque el mercado laboral les impedía adquirir una propiedad.

Según el análisis sobre el mercado español de vivienda, realizado por el observatorio de la vivienda de Century21[5], sobre la dificultad de acceso a la vivienda por los jóvenes, más del 84 % de los jóvenes españoles, entre 18 y 34 años, aspiran a tener su propia casa. A medida que mejoran las condiciones económicas y los puestos de trabajo, la proporción de personas que optan por alquilar se reduce drásticamente, hasta el 14,3 % de las personas de entre 45 y 64 años, y menos del 7% de las personas mayores de 65 años.

En términos de oferta, en 2019, más del 95 % de las propiedades en alquiler pertenecían a propietarios privados, normalmente individuos que poseían una o dos propiedades, en lugar de inversores institucionales. En términos generales, este tipo de arrendador no suele invertir lo suficiente en el mantenimiento de la propiedad, y como resultado, el mercado de alquiler se ha fragmentado profundamente y muchas casas han quedado en mal estado. También es difícil encontrar edificios completos dedicados exclusivamente al mercado de alquiler, ya que la oferta tiende a extenderse a una amplia gama de propiedades diferentes. Este panorama está impidiendo que los inversores institucionales acumulen grandes carteras y los obliga a centrarse en diseñar y desarrollar nuevos proyectos con diferentes formatos.

En otros países europeos, sin embargo, es más común encontrar propiedades en alquiler gestionadas por firmas especializadas, lo que viene determinado por las preferencias de los propietarios que, por cuestiones de conveniencia, por lo general prefieren que la gestión de su patrimonio en alquiler lo realicen profesionales.

Las Socimis comenzaron a desempeñar un papel más destacado en el mercado de alquiler, aumentando su stock de alquiler en un 57 % en 2018 y elevando su número total de apartamentos en propiedad a aproximadamente 42.000. El resto de los inversores institucionales que no cotizan en bolsa poseen aproximadamente 100.000 unidades de alquiler. Aunque estas empresas han ampliado sus carteras en los últimos años, su cuota de mercado sigue siendo relativamente baja y está muy lejos de las que poseen los grandes propietarios de propiedades que generan ingresos en Europa. Por ejemplo, en Alemania, una sola empresa, Vonovia, posee

[5] II Observatorio de la vivienda en España (2019).

más de 350.000 propiedades de alquiler, siete veces más que todas las empresas de alquiler cotizadas juntas en España. El mayor operador en España es el fondo de inversión Blackstone, que posee aproximadamente 30.000 viviendas a través de varias otras empresas (Testa, Fidere, Albirana Properties, Torbel, Euripo Properties). Los otros grandes actores del mercado son Lazora, con más de 6.000 viviendas, Vivenio, propiedad del fondo de pensiones holandés APG y la inmobiliaria Renta Corporación, con una cartera de 3.200 viviendas, y Encasa Cibeles y Tempore Properties, propiedad de Goldman Sachs y del fondo TPG RE Partners con más de 2.000 unidades cada uno.

La tendencia en términos de esfuerzo en el acceso a la vivienda, medido como el porcentaje de los ingresos que se dedican al pago de la vivienda, se encontraba en un 29 % en 2017 y se fue creciendo progresivamente hasta el 31 % en el tercer trimestre de 2019, cuando alcanza su máximo histórico. En 2020 empieza a mostrar síntomas de enfriamiento con una caída al 30,2 % (MITMA, 2022b). La OCDE (2018) muestra que los españoles destinaron uno de cada cuatro euros de sus ingresos a pagar la renta. En este mismo informe, apunta que los hogares con menos recursos, aquellos en el quintil más bajo en ingresos, el esfuerzo se dispara hasta cerca del 40 %, siendo una de las tasas más elevadas del mundo, solo por debajo de Chile, Nueva Zelanda, Grecia, Suecia y Estados Unidos.

1.2.2.4 Perspectiva actual del mercado

En general, en la actualidad se observa un mercado inmobiliario al alza después del parón temporal del mercado a mitad del 2020 debido a la pandemia (Rodríguez López, 2019). El informe presentado por el Ministerio de Transportes, correspondiente al último trimestre de 2021 (MITMA, 2022b), muestra que el número de operaciones de compraventa escrituradas crecieron anualmente un 21,4 %, con un predominio de la segunda mano. En términos de actividad constructora, tanto los visados del obra nueva como los de rehabilitación crecieron en un 32,7 % y un 6,3 %, respectivamente. Para este periodo, los precios de compraventa también siguieron en la senda del crecimiento con un 2.6 %, y la demanda internacional se mantuvo fuerte, con un total del 16,7 % de operaciones realizadas por agentes extranjeros, mayoritariamente europeos. Tanto el IPC (INE, 2022c) como los precios de oferta en los portales inmobiliarios Fotocasa e Idealista (MITMA, 2022b), muestran un ligero crecimiento de los precios en 2020 y un estancamiento el 2021. Se produjo además, un deterioro en la rentabilidad bruta del alquiler, que alcanzó su máximo en 2017 con un 4,29 % y que en 2021 se situaba en el 3,67 % (Banco de España, 2021). La intensidad del mercado inmobiliario ha mantenido su fuerza hasta el primer semestre del 2022

(Rodríguez López, 2022a, 2022b), pero el contexto de alta inflación, alza de tipos de interés e inestabilidad geopolítica pueden desencadenar procesos de ajuste futuros (Internacional, 2022).

Desde el punto de vista demográfico, el notable incremento de la población de 2,7 millones de personas entre 2007 y 2023, (INE, 2023d), de los cuales 2,2 eran de origen extranjero, ha supuesto una presión importante en la demanda del alquiler. En paralelo, se ha producido un incremento del 45% de los hogares de 1 y 2 personas, y un descenso en los hogares de 4, 5 o más personas (INE, 2021a). Ambos factores, en un contexto sin apenas creación de nueva vivienda ha conducido un re-equilibrio del mercado a través de los precios.

Consecuentemente, la accesibilidad al mercado de la vivienda residencial continua con una progresiva reducción, alcanzando los 7,6 años en 2020, el valor más alto desde 2015, y que equivale al 30,2% de la renta disponible por hogar según datos del Banco de España (2021). Estos datos contrastan con el el número de créditos hipotecarios concedidos, que crecieron un 52,62 % entre 2020 y 2021, en buena medida alentado por los bajos tipos de interés.

Por contra, el porcentaje de créditos dudosos, tanto de construcción como inmobiliarios figuran muy por debajo de las cifras del 2017, según el Banco de España (2021), con una reducción del 2,96 % en créditos para adquisición de vivienda y el 5.32 % para rehabilitación.

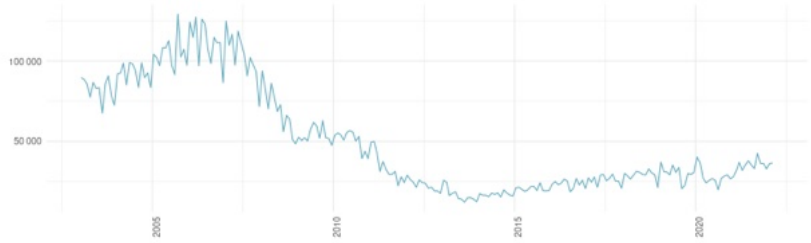
Aunque está por determinar, es posible que las situación de las rentas más bajas se haya deteriorado notablemente por la crisis sanitaria Covid-19. Si bien aún es pronto para extraer conclusiones definitivas, en el segundo semestre de 2020 se observaron cambios en los patrones de búsqueda en los portales inmobiliarios en distintos segmentos. Se atisba una polarización de la demanda: por un lado mayor interés en las viviendas nuevas; y por otro, un incremento de búsquedas de viviendas mucho más asequibles en los centros urbanos (inmuebles sin ascensor o peor estado de conservación).

Desde que la economía se hundió en la recesión de 2008, los bancos han endurecido las condiciones para acceder a la financiación hipotecaria. El “loan to value” (en adelante LTV[6]) se ha limitado al 80 % del valor de la propiedad, mientras que los requisitos mínimos de hipoteca son ahora mucho más estrictos que antes de la crisis. Más específicamente, a los posibles compradores con contratos de trabajo temporales o que no pueden cumplir con sus pagos hipotecarios mensuales con el 30-35 % de sus ganancias ahora se les niega una hipoteca. Como resultado, el número de hipotecas concedidas en 2018 fue 74,8 % menor

[6] El *“Loan To Value”* es una magnitud que representa el endeudamiento de un activo en relación con su valor real.

que en 2006 (véase Figura 1.8), con una reversión de la caída a partir del 2014.

Figura 1.8. Número de hipotecas formalizadas



Fuente: INE (2022).

Según las estadísticas catastrales, del año 2020 (Direccion General del Catastro, 2020), el dato medio en España es de 0,5 viviendas por habitante. Barcelona, Madrid, Sevilla y Las Palmas destacan por ser las provincias con menor parque inmobiliario *per cápita*, por debajo de 0,5 viviendas por habitante. En el otro extremo se encuentran las provincias de Castilla y León. Las provincias menos pobladas muestran un número mayor de viviendas, como se observa en la Tabla 1.1. Las viviendas son mayoritariamente plurifamiliares, pisos o apartamentos, que cuentan de media con una superficie de 93 m² y una antigüedad superior a 40 años.

Tabla 1.1. Máximos y mínimos provinciales en parque inmobiliario y antigüedad de vivienda

Provincia	Viviendas/Persona	Provincia	Año construcción
Ávila	1.00	Almería	1987
Soria	0.90	Melilla	1986
Teruel	0.80	Málaga	1986
Zamora	0.70	Alicante	1984
...		...	
Las Palmas	0.41	Ourense	1963
Sevilla	0.41	Zamora	1963
Madrid	0.40	Lugo	1962
Barcelona	0.40	Teruel	1959

Fuente: Catastro (2020)

Según el informe anual del mercado inmobiliario desarrollado por Idealista/data (2022), existe un proceso de densificación de población de las capitales de provincia. Las que tienen menor tasa de viviendas por habitante muestran los valores más bajos de densidad de población en el resto de la provincia, demostrando que las provincias más rurales tienen mayores diferencias de densidad de población rural y urbana.

Por ejemplo, según el INE, en torno al 22 % de la población ocupada en Canarias lleva menos de un año en el mismo puesto de trabajo. Esto está vinculado al empleo estacional, tendencia que se mantiene estable desde 2008 y tiene su impacto en el mercado del alquiler de la región. En extremo contrario, regiones como País Vasco, Extremadura, La Rioja y Castilla y León, la proporción de viviendas arrendadas se sitúa por debajo del 18 %.

Al no existir un censo oficial de oferta, se desconoce exactamente el volumen de inmuebles en comercialización. Aunque, dada la alta concentración del sector en medios de Internet, los datos del líder, Idealista, pueden servir como guía para aproximar el orden de magnitud del tamaño del mercado. Según su *Indicador de Stock Inmobiliario Provincial*[7], resumido en la Tabla 1.2, las provincias con mayor porcentaje de anuncios de viviendas en venta sobre inmuebles construidos se concentran alrededor del litoral mediterráneo. Particularmente, las valores más altos corresponden a: Málaga, Alicante y Granada, y los más bajos a: Soria, Cáceres y Teruel. La distribución en alquiler se asemeja a la de compraventa pero con mayor concentración en las provincias de Madrid y Barcelona, y situándose en el lado contrario: Cuenca, Soria y Teruel.

Tabla 1.2. Provincias con mayor y menor proporción anuncios / stock de viviendas

Compraventa		Alquiler	
Provincia	**Anuncios/Propiedades**	**Provincia**	**Anuncios/Propiedades**
Málaga	10,9%	Madrid	3,6%
Alicante	9,4%	Málaga	3,5%
Granada	8,8%	Baleares	3,5%
Baleares	8,6%	Barcelona	2,9%
Girona	6.7%	...	
...		Zamora	0,2%
Zamora	2,2%	Teruel	0,2%
Navarra	2,1%	Soria	0,2%
Cáceres	2,1%	Cuenca	0,1%
Soria	1,6%		
Teruel	1,4%		

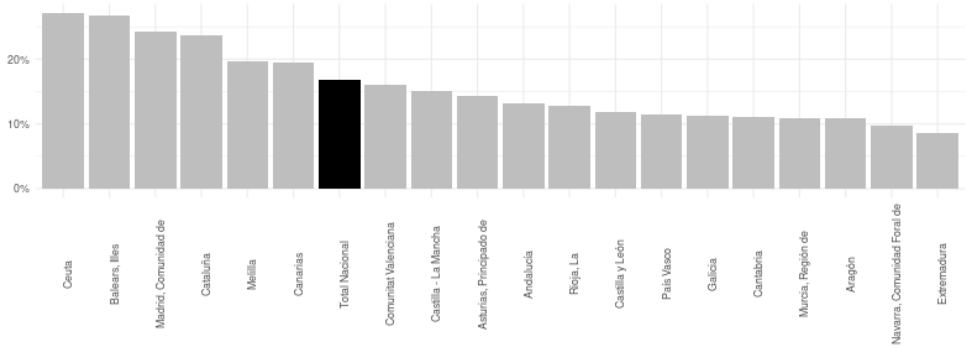
Fuente: Idealista (2020)

Las diferencias regionales mostradas en la Figura 1.9 tienen varias causas, Madrid y Cataluña, por ejemplo, son atractivas por motivos económicos y laborales y tienen mercados de alquiler dinámicos con precios altos. Las regiones insulares, junto con Ceuta y Melilla, son mercados de alquiler muy desarrollados, dado que una parte importante de los residentes planea volver a la península, evitando así

[7] Número de anuncios distintos publicados en el portal Idealista en 2020. Fuente: Idealista (2020).

vínculos adicionales con la región en el largo plazo. Echaves y Martínez (2021) encuentran también esta diversidad en la proporción de precios de compraventa y alquiler, que relacionan con la tasa de vivienda en alquiler.

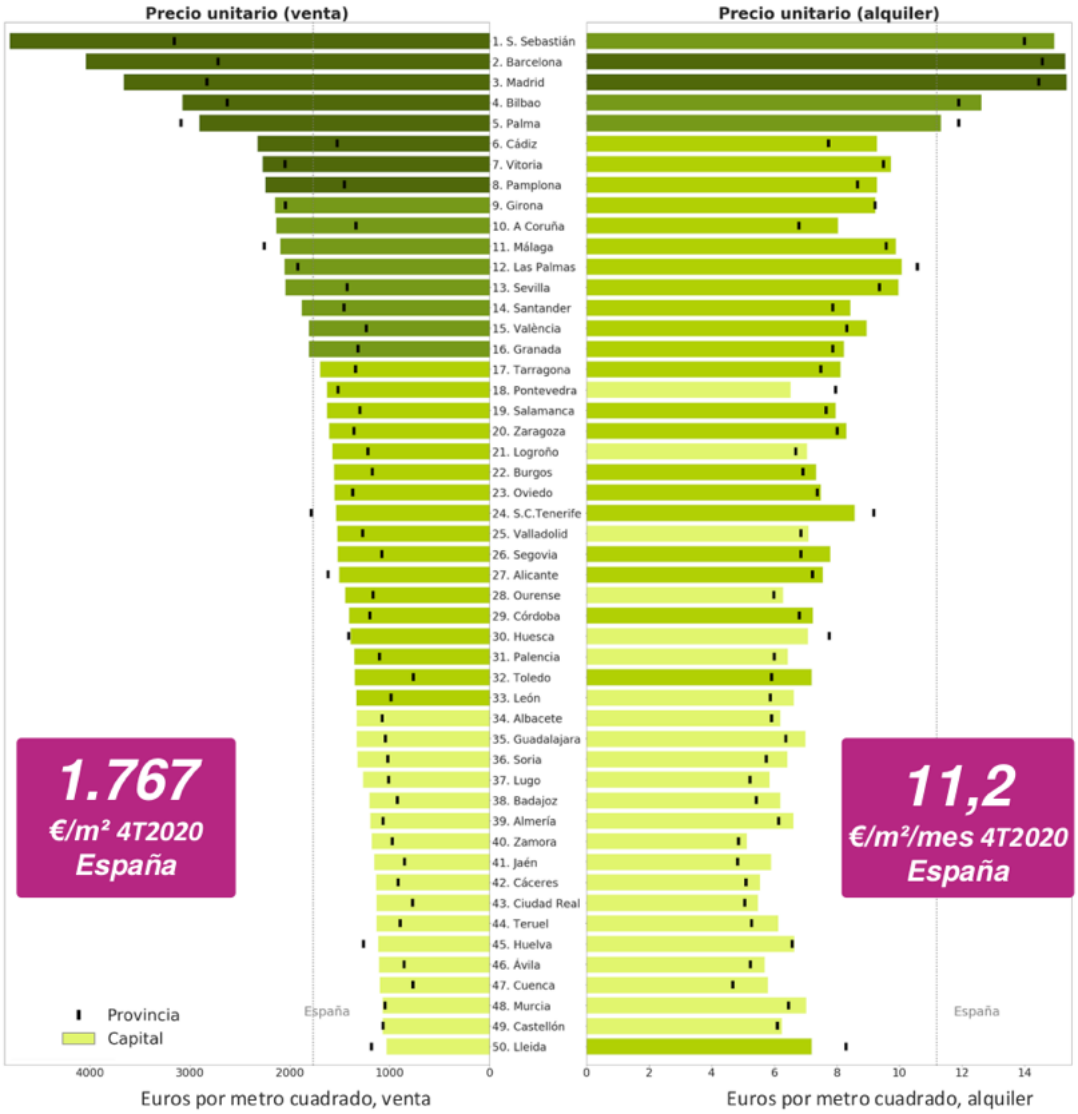
Figura 1.9. Porcentaje de hogares en régimen en alquiler



Fuente: INE (2017).

Las capitales de provincia con precios de compraventa más altos, coinciden con aquellas donde los precios del alquiler también son máximos. Como muestra la Figura 1.10, los precios unitarios €/m² de venta en las capitales son superiores al de sus provincias. Siendo mayor, en general, la diferencia en porcentaje entre capital y provincia en los precios de compraventa que en los precios de alquiler. Para ambas operaciones de compraventa y alquiler, lideran la clasificación San Sebastián, Barcelona, Madrid y Bilbao. Estas cuatro ciudades, junto con Palma, son las únicas capitales con precios unitarios de renta por encima de la media nacional, y que incluso pueden llegar a superar en un 80% los precios de las ciudades más asequibles.

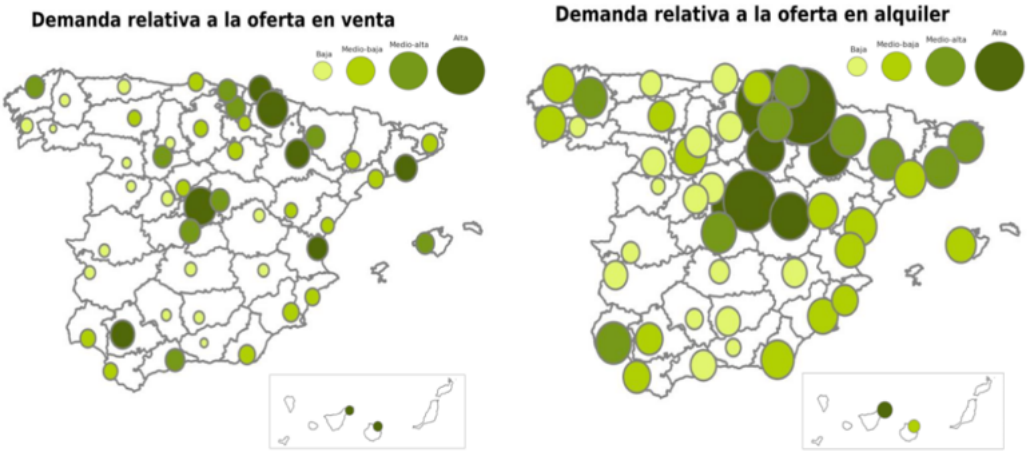
Figura 1.10. Comparativa de los precios unitarios en oferta (€/m²) para compraventa y alquiler (2020)



Fuente: Idealista (2020).

El *indicador de la demanda relativa*, medido como número de mensajes medios que recibe el propietario por cada anuncio en una zona, permite comparar la intensidad de la demanda de la oferta entre diferentes áreas, y por tanto distinguir zonas con mayor o menor interés inmobiliario. La Figura 1.11 muestra que las zonas con mayor demanda tienen precios más altos. Por lo tanto, Madrid, País Vasco, Barcelona, Sevilla y Valencia lideran el ranking de precios, y por el contrario Ourense, Granada, Zamora y Salamanca presentan las cifras más bajas. El indicador muestra valores superiores en alquiler que en compraventa, con una alta correlación entre ambas magnitudes.

Figura 1.11. Indicador de demanda relativa compraventa y alquiler (2020)



Fuente: Idealista.

En cuanto a las capitales de provincia, existe una mayor intensidad en el mercado del alquiler en términos absolutos con respecto a la compraventa. Siendo la media nacional en alquiler 5,7 puntos porcentuales, contra 1,5 puntos para el otro segmento. Entre las capitales con mayor demanda relativa se sitúan Pamplona, Vitoria-Gasteiz, Guadalajara y Zaragoza.

También destacan las zonas turísticas como Santa Cruz de Tenerife y Palma de Mallorca. El mercado de alquiler de España está creciendo más rápido que el mercado de ventas. En junio de 2019, los precios de alquiler alcanzaron un máximo histórico de 11 € por m² / mes (Idealista, 2020). Las provincias que registran las rentas más elevadas son Barcelona y Madrid, con una media superior a los 16 € / m² / mes. Le siguen Baleares y Guipúzcoa, donde los precios se acercan a los 13 € el m² / mes. En términos de variación interanual, los indicadores disponibles publicados por varios portales inmobiliarios, revelan un incremento medio anual en España de entre el 5 % y el 9 % entre 2014 y 2018 (Idealista, 2020), con incrementos que se disparan casi un 40 % en las ciudades de Barcelona y Madrid, para este periodo. Si bien, esta tendencia alcista comenzó a desacelerarse en 2018, aún se espera que los precios de alquiler en las principales ciudades sigan creciendo durante el periodo de 2020-2025, aunque a un ritmo mucho más reducido. Según Idealista (2020), se espera un nuevo repunte en Madrid, mientras que los precios se mantendrán estables en Barcelona, aunque se debe tener en cuenta que ambas ciudades registran grandes diferencias entre diferentes distritos y barrios.

Por último, es importante destacar que tanto en las ciudades de Madrid como en Barcelona, los fuertes aumentos de alquiler en los últimos años, un 44,3 % y un 60 % en términos nominales respectivamente para el periodo Enero de 2014 - Septiembre de 2022 (Idealista, 2022), han provocado un aumento importante

de las tasas de esfuerzo, con niveles que ahora superan el umbral recomendado, donde la demanda de alquiler se está desplazando hacia áreas fuera de la ciudad.

1.2.2.5 Efecto del turismo, build-to-rent en el mercado de alquiler

La principal razón por la que la gente alquila una vivienda es para vivir en ella de forma permanente. Sin embargo, en los últimos años, la demanda de alquileres turísticos de corta duración se ha disparado gracias al aumento de las plataformas colaborativas y al crecimiento del turismo. El número de turistas que visitan España desde el extranjero se ha duplicado en la última década, convirtiendo a España en el segundo destino turístico más popular del mundo, por detrás de Francia y por delante de Estados Unidos. Aunque la mayoría de los turistas extranjeros que visitan España se alojan en hoteles, cada vez más es más frecuente utilizar alojamientos de alquiler vacacional. Más específicamente, el 11,77 % de los turistas extranjeros se alojaron en este ultimo tipo de régimen en 2018, frente al 10,73 % en 2016 (INE, 2023a). Esto equivale a un aumento de 1,7 millones de personas en solo dos años, lo que eleva el total a más de 9,7 millones. Según datos de INE (2023e) a febrero de 2023, el número total de viviendas dedicadas a este fin era de 305.136, con una alta concentración de las mismas en Madrid, Barcelona y principales zonas turísticas.

El auge de los apartamentos turísticos ha estado envuelto en polémica por sus efectos económicos, sociales, medioambientales y urbanos, pero también por su potencial impacto en los mercados residencial y hotelero. Los estudios más recientes argumentan que se ha producido una sustitución del alquiler residencial de larga estancia por el vacacional, por la esperanza del propietario de obtener mayores réditos con esta modalidad (Ayouba *et al.*, 2020; Franco y Santos, 2021; Koster *et al.*, 2018). La literatura es profusa en el análisis de ciudades europeas y norteamericanas, por ejemplo, para Barcelona (Garcia-López *et al.*, 2020); para Lyon, Montpellier y París (Ayouba *et al.*, 2020); en los Estados Unidos (Koster *et al.*, 2018), (Barron *et al.*, 2021) y (Horn y Merante, 2017); y en varias ciudades de Portugal (Franco y Santos, 2021). En todos estos casos, se constata que la irrupción de estas plataformas coincide con incrementos en los precios de las rentas.

La administración es cada vez más sensible a la necesidad de encontrar una solución a los problemas asociados a los apartamentos turísticos. Sin embargo, no existe una posición consensuada ante cómo se debe regular el fenómeno; algunos sugieren introducir un límite en el número de días que un apartamento puede utilizarse para este fin; limitando estos apartamentos a determinadas áreas; haciendo obligatorio que tengan acceso independiente; o optando por

medidas que logren un mejor entendimiento entre el propietarios particulares, asociaciones y propietarios de apartamentos turísticos.

En el lado opuesto de la oferta, la falta de mercado de profesional de la vivienda ha impulsado el número de inversores del tipo "*Build to Rent*" (construir para poner en alquiler), que ha atraído de forma creciente desde 2013 a inversores institucionales y Socimis. La rentabilidad bruta del alquiler media [8] en España, se situaba en un 4,29 % en 2017, según Banco de España (2021), y se ha reducido de forma progresiva hasta 3,67 % en el tercer trimestre del 2021. Aún así, el retorno de esta inversión era mucho más atractivo que los rendimientos que ofrecen los productos de renta fija a principio de los años 2020, cuando el bono español a 10 años ofrece un 0,4 % (Banco de España, 2022a), y es superior a los rendimientos ofrecidos por los mercados prime de oficinas y locales comerciales.

1.3 Regulación y desequilibrios en el mercado del alquiler actual

Junto al escenario definidos en los anteriores epígrafes, es importante realizar una revisión breve de las cuestiones regulatorias a las que está sujeto el mercado del alquiler en España, y de forma particular a las políticas planteadas para controlar los desequilibrios existentes.

El control de rentas es una cuestión en continuo debate desde 2019, quizá por precaución ante una eventual alza inflacionista del precio de la vivienda que pueda causar procesos de expulsión en algunas zonas. Esta preocupación, ha motivado una gran cantidad de regulaciones, en muchos casos de naturaleza oportunista, que según Aruñada (2022) parten de un análisis superficial que intenta actuar sobre los síntomas (alzas en los precios), más que sobre las causas raíces de esta situación.

Una cuestión endémica del mercado inmobiliario español es su escasez de viviendas de alquiler social, y que son los promotores privados quienes asumen principalmente la función de su dinamización (Pareja-Eastaway y Sánchez-Martínez, 2017). El papel de la vivienda social se ha puesto más de relieve tras la última gran crisis inmobiliaria, debido al aumento del número de personas incapaces de acceder a una vivienda en propiedad.

Dado el enorme potencial que muestra el mercado de alquiler en el sector inmobiliario, especialmente sobre los grupos vulnerables, el límite entre lo que constituye la vivienda de alquiler privada y la vivienda de alquiler social es cada

[8] Calculada como 12 veces el precio del alquiler mensual, dividido por el precio de compraventa.

vez más difuso, principalmente porque la diferencia en precios entre los dos tipos de rentas es muy estrecha. Para el inversor privado, los riesgos de asumir esta responsabilidad han afectado negativamente a la disponibilidad de la vivienda libre en alquiler, puesto que éstos se han dedicado a atender a la demanda de quiénes deberían beneficiarse directamente de la vivienda social.

El mercado de alquiler privado ha cambiado su papel en el sistema de vivienda español convirtiéndose en un medio para la provisión de vivienda social. No solo por a la falta de alternativas por el lado de la demanda, como la disponibilidad de renta social, sino también inducido por una intervención pública mínima como proveedor de vivienda social. Esto entra en contraposición con la falta de incentivos del promotor y propietario privado para ofrecer vivienda a los hogares más desfavorecidos, por al elevado riesgo que ello supone, dado que este segmento cuenta con un mayor número de retrasos en el pago.

Las autoridades públicas han apoyado a los propietarios, directa o indirectamente, al ofrecer ayudas condicionadas a quienes ponen en alquiler su vivienda de acuerdo con criterios sociales. También a los solicitantes, cuándo cumplen determinados requisitos que los hacen candidatos por su condición de vulnerabilidad. Además, podríamos decir que este segmento del mercado inmobiliario ha estado abandonado regulatoriamente durante décadas y las leyes introducidas no han logrado un éxito considerable en la ampliación de este sector, como apreciamos en las cifras de las secciones anteriores.

En cualquier caso, esta discusión se centra realmente en el ámbito social más que en el económico, puesto que las tensiones en los precios podrían dar lugar a ciertos procesos de gentrificación o expulsión de personas que debido al aumento de los precios en una zona se ven obligadas a buscar vivienda en otras con menor coste.

1.3.1 Enfoques de regulación de precios del mercado del alquiler

En 1915 el Reino Unido introdujo la "*Increase of Rent and Mortgage Interest Act*" con el objetivo de controlar los incrementos del coste de la vivienda, que se imitó en otros puntos de Europa y del resto del mundo, siendo parte de las normativas locales en muchas ciudades: San Francisco, Boston, Cambridge, Los Ángeles, París o Berlin entre otros. En el caso de Estados Unidos, estas normas trascendieron recientemente el ámbito municipal, introduciéndose entre 2019 y 2020 normas estatales en los estados de Oregón y California.

Las normas estadounidenses, denominadas de tercera generación (Arnott,

2003), atienden principalmente a la limitación en los incrementos de precios. En cambio, las normas europeas, denominadas de segunda generación, se centran en establecer un techo de precios en ciertos mercados, entre ellas tenemos el caso de la "*Mietpreisbremse*"[9] de Alemania, iniciada en 2015 o la normativa de limitación de precios en Cataluña del 2019.

Existe una extensa literatura económica acerca de los efectos de los controles, desde Bloomberg (1947) o Rodwin (1950), que han esgrimido numerosos argumentos teóricos para resaltar las consecuencias negativas sobre el eficiencia económica de mantener las rentas por debajo de los precios de mercado. Una de ellas es el temor a introducir distorsiones entre las relaciones del mercado de compraventa y el de alquiler, el desequilibrio entre la oferta y la demanda, la ineficiencia en los consumos de servicios del hogar, o el desincentivo del propietario de mantener la vivienda en un adecuado estado, que en el largo plazo derivaría en un deterioro del vecindario (Arnott y Shevyakhova, 2014).

Como contrapartida, el control se postula como un instrumento para asegurar el mantenimiento del capital social de un barrio. Aunque existen posiciones como Olsen (1988) que cuestionan que el control de rentas tenga como efecto el empeoramiento de las condiciones de la vivienda, y que se trata principalmente de un enfoque deficiente a la hora de controlar estos elementos en los modelos econométricos usados.

Sin embargo, según Diamond, McQuade y Qian (2019), existe poca evidencia empírica bien fundamentada sobre sus efectos. Este estudio, sobre la ley que limitaba los precios del alquiler en San Francisco en 1994[10], muestra que el tope sobre las rentas redujo ligeramente sus precios en el corto plazo, pero en cambio, impactó en la movilidad de las personas de estas zonas, y se produjeron reducciones importantes en la oferta. Esta disminución fue consecuencia de la decisión de propietarios de sacar estas unidades del mercado para venderlas, al no ser atractivo el rendimiento del alquiler. En el largo plazo se observaron incrementos en los precios, que es precisamente el efecto contrario que perseguía la norma municipal. Smith and Tomlinson (1981) examinaron el control de rentas en Ontario, Canadá, encontrando un patrón similar al observado en San Francisco. La regulación en Ontario estableció un límite en los precios de la vivienda usada, que produjo una disminución en los precios nominales en este segmento, pero también tuvo un impacto significativo en la disponibilidad de la vivienda, generando un mercado dual entre las unidades sujetas a control de precios y las que no lo estaban.

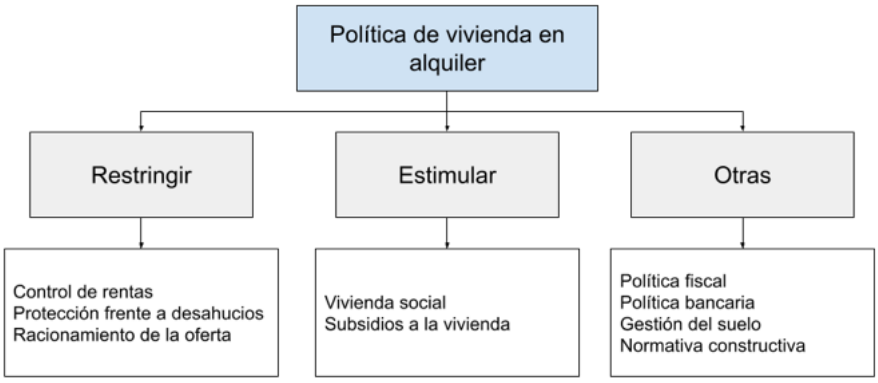
[9] Literalmente en alemán freno a los precios del alquiler.

[10] Esta normativa municipal, aprobada en 1994 en San Francisco, ampliaba el control de rentas máximas de 1979, sobre edificios de 4 o menos unidades, a todas las viviendas plurifamiliares.

La política de vivienda, en un sentido amplio, puede definirse como el conjunto de todas las medidas que aplica un gobierno para afectar el desempeño del mercado de la vivienda (Kholodilin, 2020). El objetivo principal de estas intervenciones es proporcionar a las personas una vivienda que sea asequible y, al mismo tiempo, debe satisfacer un mínimo nivel de calidad. Aparte de esto, el gobierno puede perseguir objetivos adicionales: estabilidad política, competitividad de la economía nacional e incluso estimulación de la industrialización.

Los gobiernos tienen a su disposición una gran cantidad de herramientas para regular los mercados de la vivienda. Los instrumentos de la política de vivienda, en sentido estricto, pueden clasificarse como estimulantes o restrictivos, como se describe en la Figura 1.12. Las políticas de vivienda estimulantes se presentan en dos formas: ayudas objetivas, que amplían la oferta con la construcción residencial (social); y ayudas de sujetos, que asisten a los inquilinos a través de asignaciones de vivienda. Las medidas restrictivas abarcan el control de los alquileres, la protección de los inquilinos contra el desalojo y el racionamiento o limitación de la oferta.

Figura 1.12. Políticas públicas sobre el mercado del alquiler



Fuente: elaboración propia.

1.3.2 Políticas públicas para el control del alquiler en España

En España, la creciente demanda de vivienda en alquiler en un mercado con una oferta constante, ha provocado tensiones en los precios de las rentas, por tanto, muchas administraciones han adoptado recientemente medidas para regular y crear un mercado de alquiler más accesible. Entre estas medidas, podemos citar: el incremento de la cantidad de vivienda social de alquiler, la mejora de la accesibilidad de la vivienda, ayudas a los jóvenes a salir la vivienda familiar o atraer viviendas desocupadas al mercado de alquiler (Tucat Pablo, 2021).

Existen incentivos fiscales para el inquilino, articulados como una deducción por alquiler en la renta dividida en dos tramos, siempre y cuando el alquiler se haya

formalizado antes de 2015 (Finect, 2021). El primero, estatal, que se aplica en todas las declaraciones y un segundo autonómico. Por tanto, las deducciones son distintas dependiendo del lugar de residencia. En el caso de la estatal, la desgravación es del 10 % sobre una base máxima de 9.040 euros. Sin embargo, sólo quienes sumen una base menor que 17.707 euros podrán practicarla, en caso contrario esta deducción tiene carácter progresivo de forma que la base sobre la que se practica la deducción va disminuyendo hasta desaparecer a los 24.107 euros.

Las condiciones de las deducciones autonómicas son específicas por de cada comunidad, en el caso de la Comunidad de Madrid por ejemplo, está limitada a los menores de 35 años, que pueden desgravar el 30 % del alquiler hasta un máximo de 1.000 euros. En esta misma comunidad, las personas en el tramo de edad entre 35 y 40 años, que hayan estado más de 183 días en paro en el último ejercicio fiscal, con cargas familiares podrán tener derecho a desgravación siempre que el coste de la renta supere un porcentaje de sus ingresos brutos.

El año 2013, se redujo la duración de los arrendamientos con un decreto firmado para hacerlos de tres años, aunque a lo largo de la legislatura entre 2019 al 2023 se amplió de nuevo a cuatro años.

A partir 2021, la administración propuso dos vías distintas de regulación de los precios. La primera aboga por un control del precio a través del incentivo de la oferta, a través de las reducciones fiscales: en este momento se aplica una reducción del 60 % a todos los arrendadores, por lo que el 60 % de lo que perciben por rentas de alquiler están exentos de tributación, por tanto no se suma a su base imponible en la declaración del IRPF. Por otra parte, el MITMA propone una rebaja en la bonificación con carácter general al 50 %, para incentivar a algunos propietarios si cumplen ciertas condiciones. Por ejemplo, si arriendan a personas jóvenes o ceden los pisos a programas públicos, se podría lograr una reducción del 70 %. Con ello se intenta reducir las tensiones de precios en aquellas zonas presionadas. En ellas, si un propietario realizara una reducción del 10 % la renta respecto al contrato anterior (por renovación o por renegociación con el inquilino), la bonificación llegaría hasta el 90 %. Asimismo, las viviendas que se alquilaran por primera vez al mercado de alquiler partirían de una reducción del 70%. En el caso la segunda aproximación, se aplica un tope a los precios de alquiler en las zonas tensionadas, estableciendo un precio de referencia en estas zonas.

En marzo de 2022 entró en vigor la Ley de la Vivienda que entre otras cuestiones regula los incrementos en los precios de las rentas. Hasta entonces la normativa estatal se regía por la Ley de Arrendamientos Urbanos (LAU), que se encarga de regular todos los alquileres de bienes inmuebles en España. La normativa se

había modificado por última vez en 2019, con la sanción del Real Decreto-Ley 7/2019, de 1 de marzo. A partir de este, se introdujeron condiciones más favorables para el inquilino de las que preveían las disposiciones anteriores, que databan del año 1994. Entre las medidas más importantes se encontraban: la extensión de la duración de los contratos de alquiler, el aumento de la prórroga tácita de ellos y la obligación de actualizar los valores de las rentas según el Índice de Precios de Consumo (IPC).

El cambio de la Ley de Vivienda recoge una congelación de las rentas en las áreas tensionadas para los grandes tenedores (personas físicas o jurídicas con más de 10 viviendas o más de 1.500 m^2 construidos, garajes y trasteros excluidos). Una zona se considera con tensión de precios si cumple dos criterios: el primero, si los incrementos de precios de las rentas han estado más de cinco punto por encima del IPC, durante los últimos cinco años; y el segundo, que las familias deban destinar más de un 30 % de sus ingresos para pagar el alquiler. Existen, además, incentivos fiscales para pequeños propietarios y la calificación de un 30 % de las viviendas del parque público como social, de las que la mitad se destinarían al alquiler social. Se incrementan también los impuestos sobre las viviendas vacías y se crea un bono de ayuda para el alquiler joven.

1.3.3 Políticas públicas de control de rentas otros países

Las políticas de regulación de precios máximos han alentado el debate social. Las experiencias de Italia, Estados Unidos, Alemania, o Francia apuntan dos cuestiones importantes: la primera que es algo que probado en distintas ciudades; y la segunda, que ninguno de estos casos ha logrado su objetivo de hacer la vivienda más asequible. De hecho, puede comprobarse en el estudio realizado por Kholodilin (2020), sobre la actividad regulatoria en 101 países entre los años 1910 y 2020. Las medidas de tipo restrictivo[11] han sido de aplicación común a lo largo del tiempo, particularmente en los países de nuestro entorno cercano. En la Tabla 1.3 se comparan los mecanismos de control aplicados en Alemania, Francia y España (para el que se toma la norma de Cataluña).

Alemania, con una alta densidad de población y dónde casi la mitad de las personas viven en viviendas en alquiler (un 48,9 %), registró un incremento de las rentas, entre 2010 a 2018, superior al 40 % en las siete mayores ciudades del país, según datos de la consultora Empirica (2021).

[11] Control de rentas, protección de inquilinos ante desahucio y racionamiento del mercado de vivienda.

Tabla 1.3. Método de control híbrido en España, Francia y Alemania

País	Nombre de la ley	Validez temporal	Regiones	Excepciones	Definición de los precios de referencia	Límites superior e inferior
Alemania	MietNovG de 21/04/2015	5 años, renovada en 2020	municipios y zonas asociadas	Nueva construcción / viviendas completamente renovada (gran obra) iniciadas desde 01.10.2014	1) encuesta de rentas típicas por barrio, actualizada cada 2 años (*Mietspiegel*); 2) informes de tasadores; 3) rentas de 3 domicilios; 4) base de datos de rentas	≤ 1.1
Francia	Loi ELAN de 23/11/2018	5 años	municipio de París, municipios del Grand Paris, Lyon, y Aix-Marsella-Provenza	no	mediana de rentas basadas en los precios del Observatorio Local de Rentas por categoría de vivienda y zona geográfica	[0.7,1.2]
España	Ley 11/2020 de 18/09/2020	indefinida	60 municipios de Cataluña	vivienda nueva / viviendas completamente renovada (gran obra) hasta 2023 y desde entonces durante 5 años, grandes vivienda (150+ m^2)	rentas basadas en un registro oficial (*Registro de fianzas de alquiler de fincas urbanas*)	[0.95,1.05]

Fuente: elaboración propia

El Gobierno federal intentó en 2015 reducir las subidas de precio aprobando la ley del freno del alquiler, que limita los precios en los nuevos contratos a un máximo del 10 % del alquiler medio de la zona donde está la vivienda. La norma se aplica a áreas tensionadas.

Un estudio del grupo de expertos del Instituto Alemán de Estudios Económicos (DIW), apuntó en 2018 que solo se consiguió una desaceleración moderada[12]. En zonas donde en los años precedentes los precios habían subido mucho (como los barrios berlineses de Kreuzberg o Neukölln, o el centro de Munich), el control sí evitó repuntes importantes en los precios. En enero de 2020, entró en vigor otra norma del Gobierno de la Ciudad-Estado de Berlín (a propuesta por una coalición de ecologistas, socialdemócratas, y poscomunistas) que fijaba un máximo a los alquileres en función del año de construcción del inmueble y las mejoras que hubiera tenido durante los últimos cinco años (esta medida no se aplicaba a pisos nuevos). La ley congelaba las rentas al nivel de junio de 2019 y las mantendría hasta 2025, solamente permitiendo un ligero aumento de hasta un 1,3 por ciento anual en línea con la inflación.

El instituto económico alemán IFO[13] publicó en 2022 un estudio que asegura que los alquileres habían bajado, pero la oferta se contrajo (Dolls *et al.*, 2021). Además, se observan incrementos en los segmentos no regulados y en las zonas no reguladas alrededor de la ciudad . Los propietarios sacan al mercado menos pisos, en una ciudad donde el 85 % de habitantes viven de alquiler y a la que llegan cada año 40.000 nuevos vecinos. Esta ley fue derogada en 2021 por el tribunal supremo alemán por considerarla anticonstitucional[14].

En Italia existe un mecanismo de control de rentas denominado “alquiler de tarifa acordada”, en vigor desde 1998. Este modelo permite que cada Ayuntamiento, a partir de acuerdos entre las asociaciones locales de propietarios y representantes de los inquilinos, que fije los precios máximos y mínimos de las rentas. En el caso de Roma, esta normativa se introdujo en 2004, y dispone de un método propio de valoración del inmueble. Para establecer el baremo de precios que guía las negociaciones entre arrendatario y arrendador, se tienen en cuenta incluso los servicios cercanos y áreas verdes. Este tipo de contratos, de carácter voluntario, son cada vez más populares, ya que el Estado ofrece incentivos fiscales. Se contempla un pago de tasas para los propietarios del 10% del total de la renta (frente al 21% en los denominados contratos libres).

En París, los precios de la vivienda son tan altos que más del 60 % de habitantes

[12] DIW (Evaluation of the effectiveness of the rent control introduced in 2015 to dampen rent increases in tight housing markets - Mietpreisbremse).

[13] IFO (Berlin’s Rent Cap Slows Rents and Shrinks Supply of Rental Properties, Febrero 2021. Fuente: Instituto IFO).

[14] Deutsche Welle (El supremo de Alemania frena el tope de alquileres en Berlín, 15 de abril de 2021).

de la capital francesa viven de alquiler. Aún así los precios del alquiler han experimentado una alza importante en la última década, y se estima que por encima de un 50 %[15]. Por este motivo el Gobierno francés aprobó en 2014 una ley que permitía introducir una limitación de los alquileres en las zonas denominadas como tensionadas. La justicia derogó la ley en 2017, hasta que en julio de 2019 se volvió a establecer la limitación de las rentas en la ciudad. Dicha norma se aplica tanto a contratos nuevos como a renovaciones, en ella el precio no puede ser mayor en un 20 % ni menor en un 30 % al precio de referencia anual fijado. Este precio de referencia se establece por decreto anualmente en base a la ubicación o las características de la vivienda por el Observatorio de los Alquileres de París, donde participan sector privado y administraciones. Este organismo constató que los precios se elevaron ligeramente más en 2019, cuando ya estaba en vigor la ley para contener precios, que en 2018[16]. Uno de cada tres nuevos contratos subió la renta por encima de lo que permite el índice oficial sin justificación (se podrían justificar cambios en casos de mejoras en la vivienda).

[15] Precios medios por metro cuadrado del alquiler de apartamentos y casas en París en 2021, Statista Research (2022), https://www.statista.com/statistics/769062/rent-the-metre-square-apartments-by-districts-paris-la-france/.

[16] Evolución del alquiler de viviendas del sector privado en 2018, Observatorio de los Alquileres de París (2019), https://www.observatoire-des-loyers.fr/sites/default/files/olap_documents/rapports_loyers/Rapport%20Paris%202018.pdf.

Capítulo 2

Metodología y fuentes de información

"El verdadero método de conocimiento es el experimento"

— William Blake

2.1 Introducción

La metodología propuesta pretende construir un índice de precios de la vivienda en alquiler preciso y actualizado. Esto plantea múltiples retos, el principal es la inexistencia de registros oficiales públicos de operaciones de alquiler para todo el territorio español.

Al contrario del mercado de compraventa de vivienda, donde el INE elabora el Índice de Precios de la Vivienda a partir de datos del Consejo del Notariado, en el alquiler no existe un índice equivalente. Habría que remitirse a varias fuentes: por un lado, los censos de la vivienda (realizados por el INE cada más de 10 años); y por otra parte, la Encuesta de Presupuestos Familiares (EPF) que ofrece estadísticas con cierto nivel de desglose desde el punto de vista del gasto.

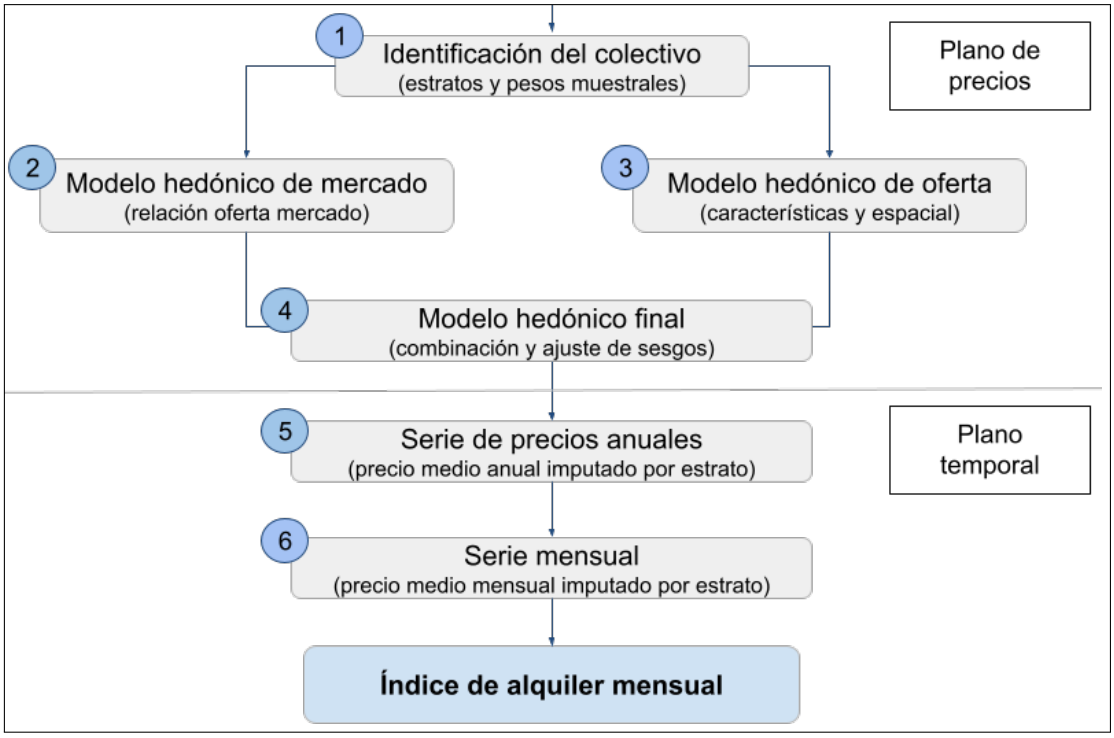
Sin embargo, en 2020, el MITMA comenzó a publicar el índice de precios del alquiler, que ofrece estadísticas estatales de precios por metro cuadrado procedentes de registros tributarios. Adicionalmente, existen algunos registros regionales y locales, como el caso de INCASOL[1] en Cataluña, que no son abiertos pero que sí se han utilizado como fuente en el ámbito investigador. Además, se dispone cada vez más de fuentes de información abierta, como el caso de los

[1] INCASOL, abreviatura de *Institut Català del Sòl*, en castellano Instituto Catalán Suelo, es una entidad creada por la Generalitat de Cataluña encargada todas las materias de urbanismo que le competen.

portales inmobiliarios, los cuales ofrecen un dato actualizado y detallado desde el ángulo de la oferta, y que varios trabajos demuestran su alta correlación con los registros oficiales (Chapelle y Eymeoud, 2022; Monràs y Montalvo, 2022).

El método se compone de dos grandes bloques que se representan en la Figura 2.1. El primero, se centra en el plano de los precios y realiza la estimación de los precios de los distintos estratos que conforman la población en alquiler (pasos del 1 al 4 en la Figura). El segundo modela el plano temporal, generando las series temporales del índice de precios de la vivienda, tanto anuales como mensuales (pasos 5 y 6).

Figura 2.1. Fases de la metodología



Fuente: elaboración propia.

El paso 1 construye los elevadores muestrales de oferta y mercado; el 2 crea un modelo hedónico de mercado que permite estimar los precios del alquiler a partir de los precios de oferta; el 3 crea un modelo hedónico de gran detalle para calcular los precios de oferta; el 4 corrige los sesgos de los modelos creados en los pasos 2 y 3, particularmente los sesgos zonales de los precios; el 5 crea las series de precios anuales de oferta y de alquiler y el 6 desagrega temporalmente los precios anuales, para finalmente construir el índice precios de la vivienda de oferta y alquiler.

La metodología tiene como objetivo el desarrollo de índices de precio del alquiler con un alto nivel de desagregación temporal, funcional y geográfico. Al no existir fuentes información con ese nivel de desglose, se utilizarán distintos modelos de

correspondencia estadística para relacionar todos los conjuntos de datos. Algunas de las fuentes utilizadas son: el Censo de Viviendas de 2011, las series de precios de la EPF, datos de portales inmobiliarios e información catastral.

Los datos de anuncios de portales inmobiliarios se utilizan para incorporar al índice un alto de nivel de desagregación funcional (características), temporal y geográfica. Debido a que los datos de oferta y alquiler guardan una fuerte relación (Chapelle y Eymeoud, 2022), motivada por el alto nivel de uso de los portales *online* en la búsqueda de vivienda.

Sin embargo, los pesos poblacionales de la oferta y del mercado no se corresponden exactamente. Entre otros motivos, porque existe un "mercado silencioso" para los portales como son las operaciones que hacen directamente los particulares, contratos de alquiler social o agencias que por diversos motivos no son clientes de los portales.

Como consecuencia de lo anterior, se plantean una serie de metas metodológicas que permitan crear los indicadores de precio, manteniendo el desglose y la coherencia con los datos de las fuentes públicas:

- Lograr una función que nos relacione el colectivo del alquiler con la población de oferta, esta función debe indicar cual es el peso que tiene cada una de las observaciones de la oferta en la población de alquiler. Dicha función nos permitirá extrapolar las magnitudes de la oferta sobre la población real en alquiler, dicho en términos estadísticos, los elevadores muestrales del colectivo de oferta. Este mecanismo debe resolver las cuestiones asociadas en estos procesos de ponderación, como son las de la infra o sobre representación de distintos segmentos de la población, o la propia falta de respuesta.
- Disponer de un mecanismo que permita extrapolar el comportamientos del mercado a partir de los censos de población y viviendas.
- Crear un modelo de hedónico de imputación de valores de alquiler a partir de datos generales de mercado desglosados. Esto permitirá traducir los precios de oferta a los precios del mercado del alquiler para una configuración de características dada: zona, habitaciones, tipo de vivienda, etcétera.
- Garantizar la coherencia de las series oficiales y calcularlas con una frecuencia mensual.

El ajuste poblacional intenta mitigar los sesgos de sobre o infrarrepresentación de ciertos segmentos en el portal. Por ejemplo, la cuota de mercado de un portal puede variar en función de la zona y el tipo de inmueble, y no tiene por

qué corresponderse a la distribución de las transacciones que formalizadas o al número de hogares actualmente en alquiler. Otro fenómeno habitual que introduce distorsiones poblacionales es la existencia de múltiples anuncios por cada vivienda, y que se produce cuando varias agencias y el propietario comercializan simultáneamente la misma vivienda.

Los sesgos poblacionales y de no respuesta se resuelven mediante un proceso de estratificación y de cálculo de sus pesos poblacionales. En nuestro caso, se realiza un proceso calibración de los elevadores muestrales para relacionar, en el tiempo, la población de oferta con el colectivo a modelar (el mercado de alquiler). Lo que da lugar a un colectivo de oferta altamente desglosado que mantiene las proporciones de la población del mercado.

Aunque el precio de oferta guarda relación con el precio de mercado, éstas son magnitudes distintas (Shimizu *et al.*, 2016). Para vincularla, se desarrollará un modelo que relaciona los precios de puja (el precio pedido por el propietario en el portal) y el precio negociado (el que finalmente se acuerda).

Además, la construcción del índice requiere trabajar con unidades estables y comparables en el tiempo, lo que es prácticamente imposible en el mercado inmobiliario, donde cada vivienda es única. Para lograr el equivalente a una "cesta de viviendas" sobre la que medir la evolución de los precios, se construye un modelo denominado hedónico, que permite asignar los precios de la cesta a lo largo del tiempo.

Existe un último reto asociado a la frecuencia de la información, ya que se parte de precios de mercado anuales y de precios de oferta mensuales. Como el objetivo final es disponer de índices de precios del alquiler mensual, será necesario realizar un proceso de desagregación temporal de las series anuales para convertirlas a series mensuales.

El marco teórico en el que se encuadra el trabajo de investigación es amplio y se completará en cada uno de los capítulos que desarrollan la metodología. Debido a esta amplitud de temáticas y para facilitar una visión general de las cuestiones involucradas, se muestran los aspectos teóricos soportan cada aspecto en la Tabla 2.1.

Tabla 2.1. Aspectos cubiertos por el marco teórico

Aspecto teórico	**Aplicación en la metodología**	**Plano**
Muestreo, reponderación y calibración de muestra	Construcción del colectivo y cálculo de los elevadores muestrales sobre la población de oferta	Precios
Modelos de valoración por precios hedónicos	Creación de modelos hedónicos de precios de alquiler	Precios
	Creación de modelos hedónicos de precios de oferta	Precios
	Creación de modelo de enlace entra precios de oferta y de alquiler	Precios
Econometría espacial	Modelización del componente de utilidad de la ubicación en los modelos hedónicos de oferta	Precios
Teoría de índices de precios y métodos para la construcción de índices de la vivienda	Construcción de índices base, índices encadenados anuales e índices encadenados mensuales	Temporal
Modelos de reconciliación y desagregación de series temporales	Desagregación temporal de series de precios anuales de alquiler a series de precios mensuales	Temporal

Fuente: elaboración propia

El capítulo se desarrolla en tres partes, la primera describe la aproximación metodológica del plano de los precios, con los aspectos esenciales de su marco teórico; la segunda describe el aspecto temporal de la metodología, ahondando en los métodos de construcción de índices; y finalmente, la tercera parte, muestra con detalle cada una de las fuentes de información utilizadas en el trabajo, con una exposición de los procesos aplicados para el control de calidad y corrección de las bases de datos utilizadas.

2.2 Plano de precios: Modelos hedónicos

Desde un punto de vista conceptual, la teoría de precios hedónicos se aplica a bienes heterogéneos y descansa sobre la hipótesis de que el precio de mercado de un bien complejo (Z) es función directa de su utilidad o beneficio, derivado de la cantidad de los n atributos que lo componen.

El precio de mercado de Z es el resultado del equilibrio entre la oferta y la demanda según características conocidas. Cada consumidor, o comprador, cuenta con una función de puja θ que representa su disposición a pagar por el bien, luego θ es una función asociada al bien Z, expresada como función de las cantidades individuales de sus atributos n y de la utilidad derivada (ν) para cierto nivel de ingresos (y), dada una estructura de preferencias (α). Como indica Malpezzi (2003), el modelo se deriva de la heterogeneidad del stock inmobiliario y de las preferencias de los consumidores, ya que el inmueble tiene características únicas y cada consumidor las puede valorar de manera diferente.

La función de puja se puede escribir como:

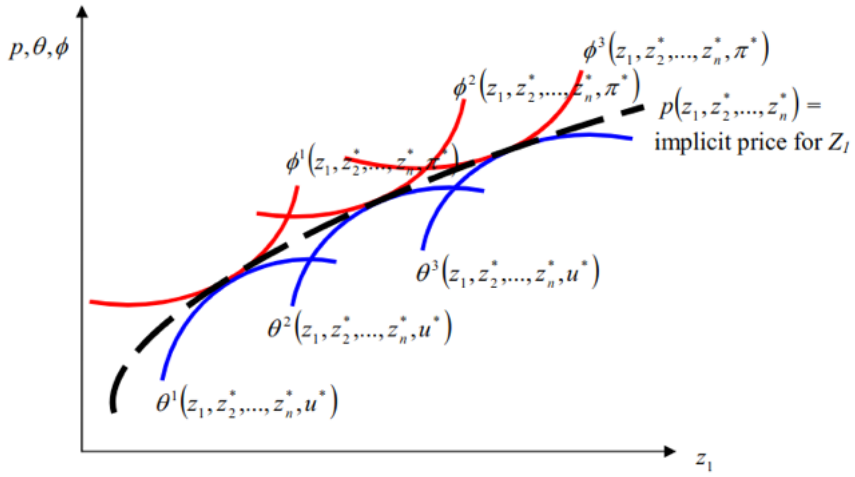
$$\theta = \theta(z_1, ..., z_n, \nu_{y,\alpha}) \qquad [2.1]$$

De forma similar, la función de oferta (ϕ) se define como el precio mínimo que el vendedor está dispuesto a aceptar por Z, considerando sus atributos y un beneficio esperado (π), para un nivel de producción (M) y una función de coste (β). La función de oferta se especificaría como sigue:

$$\phi = \phi(z_1, ..., z_n, \pi_{M,\beta}) \qquad [2.2]$$

El equilibrio de mercado se alcanza con cada atributo en el punto de tangencia entre las funciones de puja y de oferta. La Figura 2.2 muestra esta cuestión de forma gráfica para un único atributo. Se mantienen constantes todas las dimensiones distintas al atributo Z_1, mostrando como línea punteada la curva que representa la función de precios hedónicos para el atributo. La generalización de este esquema conduce a una familia de funciones (precios hedónicos) donde se alcanza el equilibrio de mercado para los n atributos del bien.

Figura 2.2. Determinación del precio implícito para el atributo z1



Fuente: Des Rosiers y Thériault (2006).

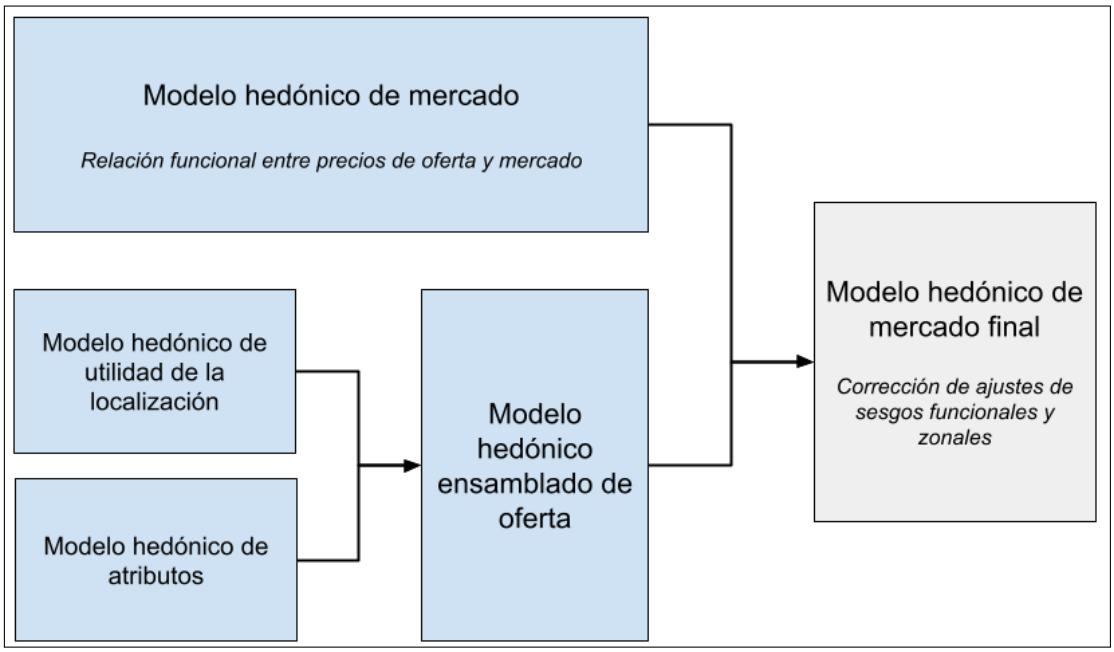
Según Rosen (1974), el precio hedónico de un bien se define como "el método mediante el cual se calculan los precios implícitos de los atributos o características que componen a un bien compuesto". Sobre esta base, Witte (1979) aplica la teoría de precios hedónicos para la valoración de precios de la vivienda en propiedad, y Thibodeau (1995) la implementa al alquiler.

Este método no es exclusivo del mercado de la vivienda residencial, sino que se ha aplicado a diversos tipos de bienes como los automóviles y ordenadores (Berndt, 1991), cambios en la valoración de suelos (Cheshire y Sheppard, 1995), y la renta implícita en la posesión de una vivienda (Gasparini y Sosa Escudero, 1999), entre otros casos.

El propio Rosen (1974) menciona dos limitaciones importantes en el marco teórico de los precios hedónicos aplicados a la vivienda. La primera, que la función mezcla de forma indiferenciada factores de oferta y demanda, por tanto, induciendo un problema de identificación. La segunda, que la linealidad de la función es cuestionable, a la luz de la evidencia empírica del comportamiento no lineal de varios fenómenos responsables de la formación de los precios inmobiliarios como: la contribución marginal del área útil al precio, que se conoce que es decreciente; la heterogeneidad espacial, y otras debidas a variables omitidas (Des Rosiers y Thériault, 2006).

En nuestro caso, para mitigar las limitaciones mencionadas, se ha desarrollado un modelo hedónico compuesto a su vez de una serie de modelos, que se muestran en la Figura 2.3.

Figura 2.3. Componentes del conjunto de modelado hedónico



Fuente: elaboración propia.

El modelo hedónico de oferta permite estimar el precio que tendría una vivienda en oferta, y se construye mediante la agregación de otros dos modelos: el de atributos y el de utilidad de la localización, cada uno especializado en un aspecto fundamental. Finalmente, se conectan los dos modelos de oferta y mercado a través de un modelo final, que corrige los sesgos incurridos en los modelos previos.

2.2.1 Tipos de modelos hedónicos

La estimación de un modelo de precios hedónico puede realizarse a través de tres enfoques: paramétricos, semiparamétricos, no paramétricos y de aprendizaje estadístico[2]. Las aproximaciones no paramétricas surgen como solución las debilidades de los paramétricas, como las no linealidades, el control de la heterocedasticidad, la heterogeneidad espacial entre otras cuestiones. A continuación se describen los cuatro tipos de regresión hedónica.

2.2.1.1 Métodos paramétricos

El enfoque paramétrico asume que existe una curva de regresión que sigue una forma funcional, especificada mediante a un número finito y conocido de parámetros. Los parámetros son, por lo general, coeficientes de variables independientes (Horowitz y Lee, 2002), y representan las contribuciones

[2] En realidad, los métodos basados en aprendizaje son de tipo no paramétrico, pero se ha distinguido en una categoría propia dado su creciente popularidad en el ámbito industrial y econométrico.

marginales de cada atributo que tiene una vivienda.

El primero, y más conocido, es la regresión lineal por mínimos cuadrados, que estima el precio a través de las contribuciones aditivas de sus características. No existe una forma funcional general para especificar el modelo, y es finalmente el investigador el que determina la mejor forma a aplicar a su caso (Owusu-Ansah, 2011). Existen tres formas principales: lineal [2.3], semilogarítimica [2.4] y logarítmica [2.5].

$$p = \beta_0 + \sum_{i=1,n}^{N} \beta_i \cdot C_i + \epsilon \qquad [2.3]$$

donde p se refiere al precio a predecir, C_i es el atributo i dentro de una lista de n atributos que describen a la vivienda, β_0 es la intersección de la regresión, los β_i los coeficientes (contribuciones) de cada de los atributos y ϵ es un término que representa el error aleatorio.

El principal inconveniente de la forma lineal es que las viviendas son bienes heterogéneos, en los que las relaciones entre las covariables y el precio no tienen porque guardar una relación lineal. Por tanto, es común encontrar fenómenos de heterocedasticidad y no linealidad que dificultan su uso en la práctica. Goodman y Thibodeau (1995) comprobaron que la relación entre el precio y las variables de entrada no es necesariamente lineal, y descubrieron que a través una forma lineal semilogarítmica se mejoraba el ajuste.

Dentro del ámbito inmobiliario, existen numerosas relaciones no lineales entre covariable y precio, como por ejemplo las variaciones en los precios de los alquileres (Thibodeau, 1995), los cambios en la valoración de suelos (Cheshire y Sheppard, 1995) y la renta implícita en la posesión de una vivienda (Gasparini y Sosa Escudero, 1999). Esta aproximación semilogarítmica es muy habitual en la literatura (Sirmans *et al.*, 2005), y cuenta con la ventaja de la fácil interpretabilidad de sus coeficientes, además de reducir la heterocedasticidad del modelo (Follain y Malpezzi, 1980).

$$log(p) = \beta_0 + \sum_{i=1,n}^{N} \beta_i \cdot C_i + \epsilon \qquad [2.4]$$

La forma logarítmica es similar a la anterior, con la diferencia de que las covariables de la expresión son el logaritmo de los atributos. Aún cuando se aplique esta especificación, en ocasiones no será una logarítmica pura, ya que si algunas de las características tienen valores cero o se utilizan variables ficticias dicotómicas para capturar la presencia o ausencia de una característica, no sería

posible aplicar logaritmos a estas variables[3] y, por tanto, dichas variables se modelan en forma semilogarítmica (Bover y Velilla, 2001).

$$log(p) = \beta_0 + \sum_{i=1,n}^{N} \beta_i \cdot log(C_i) + \epsilon \qquad [2.5]$$

Sin embargo, las formas de tipo logarítmico tienen como inconveniente que las medidas de error de los modelos ofrecen una visión distorsionada, desde un punto de vista estadístico. Por tanto, tal y como sugiere Pérez-Rave et al. (2019), las medidas de error se deberían calcular siempre en términos monetarios.

Desde Rosen (1974) hasta los trabajos más recientes, como el de Diewert (2003), se han llevado a cabo distintos estudios teóricos para determinar la forma funcional óptima en los métodos lineales. Dada la dificultad a la hora de establecer la forma funcional y la variable objetivo a modelar, Diewert (2003) sugiere un conjunto de recomendaciones para que las agencias estadísticas puedan abordar de manera organizada esta cuestión. Algunas de estas guías incluyen la decisión de si es mejor transformar la variable dependiente; si es preferible realizar una sola regresión hedónica para todos los períodos o una para cada período; si deben imponerse restricciones en los signos de los coeficientes; si deben usarse regresiones ponderadas; o cómo deben tratarse los valores atípicos.

Las covariables de la regresión son de diferente naturaleza, en ellas se incorporan tanto atributos físicos de la vivienda, como variables ficticias (*dummy*) que representan el momento en tiempo o la zona en la que se encuentra la vivienda. Se podría especificar el modelo lineal de una forma mucho más detallada mediante la siguiente expresión analítica:

$$p^t = \beta_0 + \sum \delta \cdot D_{nk}^t + \sum \beta \cdot S_{nk}^t + \sum \gamma \cdot L_{nk}^t + \sum \mu^t M_{nk} + \epsilon \qquad [2.6]$$

donde D_{nk}^t son variables ficticias dicótomicas que representan el tiempo[4]; atributos de estructura S_{nk}^t, representado por variables continuas; variables dicotómicas asociadas a la ubicación L[5]; y las características de mercado inmobiliario M_{nk}^t[6]. Estas últimas rara vez se agregan a los modelos, debido a la dificultad para obtener esta información (particularmente en los casos que usan transacciones formalizadas), sin embargo, se vuelven muy relevantes para comprender el comportamiento de las fuerzas de oferta y demanda para cada

[3] Cuando la variable toma el valor cero.

[4] Existiría una variable que puede valer uno o cero para cada uno de los periodos de tiempo que existan en la muestra.

[5] Estas variables serían tantas como zonas se consideren en el modelo, un valor 1 para una variable asociada a una zona Z, representa que la observación se ubica en la zona Z.

[6] Por ejemplo si el bien se vende un particular o una agencia inmobiliaria.

submercado (Piazzesi *et al.*, 2015).

La crítica principal a los modelos lineales de mínimos cuadrados es su difícil control de la heterocedasticidad, ya que para aplicarla correctamente debe existir homocedasticidad en el error no observado, es decir, que el error condicionado a las variables independientes debe tener una varianza constante, y que las covariables deben ser independientes para ser interpretables. Existen diferentes estudios que muestran que rara vez estas condiciones se cumplen, principalmente debido a comportamientos diversos del error en los distintos segmentos de la población (Stevenson, 2004), bien por cuestiones funcionales, de mercado o espaciales. No obstante, aun cuando la heterocedasticidad está presente, los modelos de regresión por mínimos cuadrados (MCO) son insegados pero consistentes (Fletcher *et al.*, 2000), aunque este fenómeno dificulta la interpretación de los coeficientes, ya que la heterocedasticidad afecta a la estimación de los errores estándar y magnitud de los coeficientes (Stevenson, 2004). Otra crítica a estos métodos es que comportan restricciones implícitas, entre las que se encuentran un número limitado de parámetros (Härdle y Linton, 1994).

Existe una variación sobre el modelo de mínimos cuadrados, que es el método de mínimos cuadrados ponderados (WLS[7]). Al contrario del método base, que asume una varianza del error igual en toda la población, el método ponderado ajusta el peso de las distintas observaciones de la muestra. Por tanto, se atribuye un mayor peso a las instancias con menor varianza de su error.

2.2.1.2 Métodos semiparamétricos

Las regresiones de tipo semiparamétrico, introducen información paramétrica a una regresión no paramétrica para aprovechar las ventajas de cada tipo de modelo, y reducir sus correspondientes desventajas. Partiendo de la idea de Robinson (1988), Stock (1989) aplica esta aproximación para estimar el impacto en los precios de la vivienda de eliminar material contaminante cercano al vecindario. Aparte de este primer modelos, denominado de “Robinson-Stock”, se pueden encontrar el de “Yatchew de diferencias” (Yatchew, 1997) y regresiones locales de Clapp (2004).

Otro enfoque semiparamétrico es el de los modelos generalizados aditivos (GAM), que son métodos lineales donde los coeficientes de los predictores no son valores constantes sino que se calculan a través de una función. Estas funciones, denominadas funciones base, son habitualmente curvas de suavizado (*spline*) que toman distintos valores en función del predictor. Tienen la ventaja de controlar

[7] Del inglés Weighted Least Squares.

eficazmente las no-linealidades y la heterocedasticidad (Hastie y Tibshirani, 2017), manteniendo la interpretabilidad de los coeficientes.

Existen varios casos de aplicación se GAM en el modelado hedónico, como Pace (1998), Munger (2021), Ulbl (2021) o Bax (2021).

2.2.1.3 Métodos no paramétricos

La aproximación no paramétrica no exige que la relación entre las variables dependientes e independientes sea conforme a una función de regresión (Fox, 2000). Existen múltiples técnicas, entre las que podemos encontrar: los métodos basados en splines[8] (Reinsch, 1967); el método basado en los vecinos más cercanos (kNN[9]) (Fix y Hodges, 1989; Li, 1984); métodos basados en *kernels* (Watson, 1964); o métodos basados en regresión local ponderada (LWR) o regresión local polinómica (LPR) (Cleveland *et al.*, 1988; Cleveland y Devlin, 1988).

Es importante destacar que el método de los K vecinos más cercanos es de uso habitual en los procesos de tasación inmobiliaria, y se denomina como "valoración por comparables" o "valoración por testigos". Está actividad está regulada en España por la normativa ECO/805/2003[10].

Estos modelos tienen la ventaja de no tener que ajustarse a una única forma funcional, pueden funcionar bien en casos con poca muestra (por ejemplo las regresiones locales) y/o ante valores ausentes. Son menos exigentes que los métodos paramétricos en cuanto a condiciones a cumplir por las variables de entrada, y pueden trabajar con una mayor diversidad de variables.

Una de las críticas a estos métodos es que habitualmente sufren de la denominada "maldición de la dimensionalidad"[11] cuando hay un gran número de variables (Van Der Maaten *et al.*, 2009; Zhu y Bradic, 2017). En otros casos, los basados en regresiones locales, las muestras no tienen por qué distribuirse de forma equitativa, por tanto se pueden generalizar mal en segmentos donde hay poca información o esta muy desbalanceada (Taylor y Einbeck, 2013). Una ultima desventaja es que al no haber parámetros que describan la regresión, es más complicado establecer comparaciones cuantitativas entre dos o más poblaciones.

[8] Una spline es una curva diferenciable definida en porciones mediante polinomio, en un espacio bidimensional se podría utilizar para estimar una curva a una secuencia de puntos, especificada funcionalmente con la forma de un polinomio.

[9] K vecinos más cercanos, o K "*nearest neighbors*".

[10] Más detalles en https://www.boe.es/buscar/doc.php?id=BOE-A-2003-7253

[11] También conocido como efecto Hughes, se refiere a los diversos fenómenos que surgen al analizar y organizar datos cuando el número de variables es muy alto, principalmente derivados del incremento de la complejidad de cálculo en una escala potencial o exponencial.

2.2.1.4 Métodos de aprendizaje estadístico

El aprendizaje automático, estadístico[12] o aprendizaje de máquinas[13], es un campo compartido de la estadística y las ciencias de la computación . El aprendizaje se desarrolla a través de un proceso repetitivo en el que el modelo se crea mediante de la generalización de un conjunto de ejemplos .

El aprendizaje automático está experimentando una explosión en la presente década, y su aplicación está impactando a distintos campos de la ciencia y la industria. Estas técnicas están relacionadas con el denominado fenómeno del "*Big Data*" (Demchenko *et al.*, 2014), en el que la presencia abundante y detallada de información permite construir sistemas de decisión automáticos con una alta precisión.

El sector inmobiliario no ha sido ajeno a esta evolución y los modelos de valoración sobre aprendizaje automático empezaron a aplicarse en la década de los 90 del siglo XX. Los primeros modelos en aplicarse se basaban en redes neuronales artificiales (Curry *et al.*, 2002; Ge *et al.*, 2003; Kauko *et al.*, 2002; Liu *et al.*, 2006; McCluskey y Anand, 1999; Pace, 1995; Selim, 2009; Verikas *et al.*, 2002; Worzala *et al.*, 1995).

La aproximación de estos modelos no es muy diferente de la estadística inferencial, ampliamente aplicada en el campo de la econometría, puesto que ambas disciplinas se basan en el análisis de datos. La vertiente computacional de estas técnicas enfatiza el correcto manejo de la complejidad computacional de los algoritmos[14], dado que buena parte de los problemas tienen una dificultad de cálculo no polinómica (problemas NP-Hard).

Los algoritmos de aprendizaje automático se pueden clasificar en cuatro tipos: supervisados, no supervisados, semisupervisados y de aprendizaje por refuerzo.

Los supervisados construyen una relación entre las salidas deseadas (etiquetas) y las entradas. Por su parte, los no supervisados no cuentan con una variable de respuesta concreta y crea un modelo sobre los patrones observados en los datos, un ejemplo es la detección de grupos (clustering), la detección de anomalías o la reducción de dimensiones.

Los métodos semi-supervisados son un híbrido entre métodos supervisados y no supervisados. Cuentan con datos tanto etiquetados como no etiquetados, aunque el número de registros etiquetados en sensiblemente menor que los que no lo

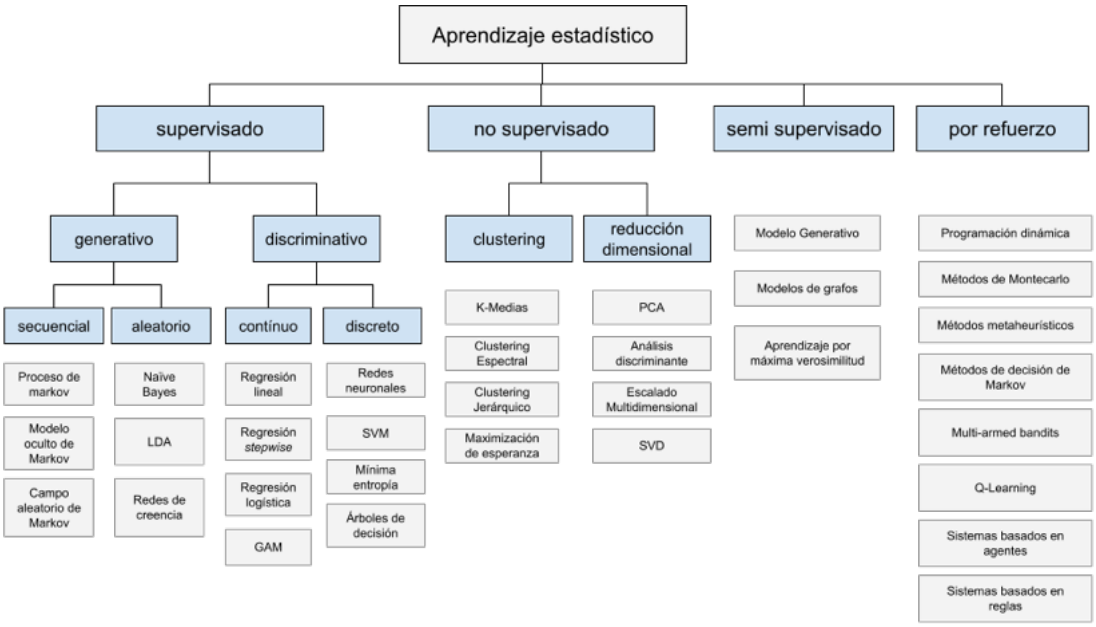
[12] Dependiendo del tipo de técnica aplicada, en particular aquellas en las que existe base estadística en el proceso, algunos autores lo califican como aprendizaje estadístico.
[13] También referido en su original en inglés "*machine learning*".
[14] La complejidad computacional se puede definir como el nivel de exigencia de recursos: proceso, memoria o tiempo, necesarios para resolver un problema con un programa de ordenador.

están.

Los métodos de aprendizaje por refuerzo se construyen mediante un ajuste continuo, a través de un proceso de estímulo-respuesta, reforzando los comportamientos más productivos y desechando aquellos que tienen peor rendimiento.

El caso de la predicción del precio de la vivienda se ajusta al tipo supervisado, al conocerse la magnitud a predecir (precio) para cada uno de los registros. En la Figura 2.4 se muestra una taxonomía de los algoritmos principales de cada familia de métodos.

Figura 2.4. Taxonomía de métodos de aprendizaje automático



Fuente: elaboración propia.

Dentro de los distintos métodos más utilizados para la estimación de los precios de la vivienda se encuentran los métodos basados en redes neuronales, K-Vecinos, SVM, algoritmos genéticos y árboles de regresión (Valier, 2020).

Las máquinas de soporte vectorial de regresión (SVM) proporcionan una única solución óptima al problema y son capaces de trabajar con muestras de datos pequeñas (Zulkifley *et al.*, 2020). Además, no requieren una distribución de probabilidad determinada ni la existencia de una relación lineal entre variables dependientes e independientes.

El primer modelo de redes neuronales artificiales (RNA o en inglés ANN[15]) fue propuesto por Pitts y McCulloch (1943) en el artículo titulado *"Un cálculo lógico de*

[15] Red Neuronal Artificial o Artificial Neural Network.

ideas inmanentes en la actividad nerviosa". Este modelo, de inspiración biológica, es adecuado para modelar relaciones no lineales complejas, y es especialmente interesante en el caso de la vivienda dado la numerosa presencia de relaciones de este tipo (Krogh, 2008). La aplicación práctica de las ANN se fundamenta en la propiedad teórica de "aproximación universal", descrita con detalle por Hornick *et al.* (1989), y que significa que las redes son capaces de adaptarse para aproximar cualquier forma funcional desconocida para cualquier grado de precisión deseada. Este concepto hace que sea posible considerar este tipo de modelos como métodos estadísticos flexibles no lineales (Curry *et al.*, 2002).

Aunque existe un consenso general de que este tipo de métodos mejoran de manera notable la precisión de las regresiones hedónicas paramétricas, Valier (2020) apunta que no son superiores en términos de inferencia, lo que dificulta su uso a la hora de extraer conclusiones de ellos. Nguyen y Cripps (2001) muestran que las redes neuronales son eficaces en conjuntos de datos grandes y heterogéneos. Otros autores constatan la misma eficacia pero utilizando otros métodos: K vecinos más cercanos (kNN) (McCluskey y Anand, 1999); técnicas de lógica borrosa (Bagnoli y Smith, 1998; Thériault *et al.*, 2005); árboles de regresión simples (Fan *et al.*, 2006), y árboles de regresión ensamblados (Baldominos *et al.*, 2018; Hjort *et al.*, 2022; Hong *et al.*, 2020).

Finalmente, ya en las primeras aplicaciones, algunos autores identificaban potenciales problemas con la falta de homogeneidad de los métodos. Por ejemplo, el error absoluto varía significativamente en función del paquete de software que se había utilizado para estimar el modelo (Kontrimas y Verikas, 2011; Worzala *et al.*, 1995).

2.2.1.5 Modelos hedónicos geográficos

La influencia de la localización en los precios, denominada dependencia espacial (Anselin y Rey, 2014), es un aspecto clave a especificar en los modelos (Hill, 2013). Particularmente el control del fenómeno de la heterogeneidad espacial (Anselin y Griffith, 1988), que hace referencia a la variación en características o atributos de una región o espacio geográfico, es fundamental para evitar problemas de especificación. La ausencia de un tratamiento apropiado puede provocar una serie de inconvenientes:

1. Especificación incorrecta del modelo: debido a la omisión de variables relevantes o la inclusión de variables irrelevantes que compensan la ausencia de especificación espacial. Cualquiera de estos dos casos puede producir estimaciones sesgadas o resultados inconsistentes debidos.

2. Predicciones inexactas: particularmente en áreas con variación significativa en factores como la calidad del vecindario, acceso a serivios e influencia de condiciones ambientales. El efecto son estimaciones erróneas de la relación entre las variables explicativas y los precios de las viviendas, lo que da lugar a interpretaciones erróneas del modelo.

3. Autocorrelación espacial: la ausencia de control de las diferencias de precios en el espacio por el modelo puede resultar en autocorrelación espacial en la variable dependiente y/o en las variables explicativas. Esta condición es especialmente grave en los modelos de regresión ordinarios, puesto que viola el supuesto de independencia de las observaciones. Las consecuencia son estimaciones de parámetros sesgadas e ineficientes que producen inferencias incorrectas.

4. Efectos de desbordamiento espacial (*spillover*): pasar por alto los efectos de desbordamiento espacial, es decir, que los cambios en una ubicación pueden afectar los precios de las viviendas en ubicaciones vecinas[16]. Cuando se ignoran los efectos por desbordamiento se obtienen parámetros sesgados o engañosos.

5. Sesgo de agregación: se produce cuando se recopilan y analizan en diferentes escalas espaciales (por ejemplo, nivel de vecindario, ciudad o región), y está asociado con el problema del área modificable MAUP (Wong, 2004). La relación entre los precios de las viviendas y sus determinantes puede variar en diferentes escalas espaciales, al ignorar esta cuestión se pueden producir efectos inconsistentes de los factores a distintos niveles (por ejemplo el impacto de un factor es más pronunciado a nivel de ciudad que en sus barrios).

6. Efectos de borde: ocurren cuando los límites del área de estudio influyen artificialmente en las relaciones estimadas entre variables, y podría estar relacionado con el punto anterior. En efecto que produce son estimaciones sesgadas en torno a los límites entre áreas.

7. Generalización limitada: la heterogeneidad espacial puede limitar la aplicabilidad y generalización de un modelo de precios de vivienda a otras regiones o períodos de tiempo. Se produce al construir el modelo en función de las características específicas de un área en particular, cuyo comportamiento no es extrolable a otras áreas con patrones urbanos y de mercado diferentes. Esta cuestión puede ser especialmente problemática

[16] Un ejemplo de influencia en los precios por desbordamiento sería la construcción de una nueva estación de transporte público en un área, lo que produciría a un aumento de los precios de las viviendas no solo en las inmediaciones, sino también en áreas vecinas debido a la mejora en la accesibilidad general de la zona.

ante la toma decisiones en diferentes áreas o a lo largo del tiempo.

8. Multicolinealidad: en ocasiones las características específicas de una localización implica altas correlaciones entre variables explicativas, lo que resulta en multicolinealidad. En el caso de regresiones ordinarias puede implicar una estimación sesgada y engañosa de los coeficientes, en otros métodos dificulta la interpretabilidad de la influencia de las variables en el precio.
9. Desafíos computacionales: el control de la heterogeneidad requiere el uso de técnicas de modelado complejas, como modelos econométricos espaciales o de panel espacial.

Los modelos hedónicos geoespaciales utilizan información geográfica, como la ubicación o distancias a puntos de interés, para mejorar la estimación del valor la vivienda resolviendo los problemas enumerados anteriormente. Estos modelos se apoyan en técnicas como la econometría espacial y la geografía cuantitativa (Anselin, 2002; Can, 1992), y se describirán con más detalle en el Capítulo 6.

Por otra parte, muchos de las aproximaciones paramétricas de la estadística espacial expresan el espacio de trabajo a través unidades espaciales definidas exógenamente (barrios, distritos, regiones), sin embargo los cambios en el espacio se producen de forma continua. Por tanto, tanto la definición de las variables geográficas como los modelos de valoración deben especificarse sobre un espacio continuo (Helbich *et al.*, 2014). A este último respecto corresponden los modelos de regresión polinomial y de expansión espacial, o los de regresión local ponderada.

En la Tabla 2.2 se muestra una taxonomía reducida de los métodos de regresión hedónica espacial puede estructurarse en función de las técnicas y enfoques utilizados, entre otros, podemos destacar: modelos basados en variables espaciales; de interacción espacial (rezagos y correlación de errores espaciales); de control de la heterogeneidad espacial; para abordar la dependencia espacial y la heterogeneidad espacial en los modelos.

Tipo de modelo	**Subtipo**	**Descripción**
Variables espaciales	Variables de distancia	Incorporan distancias a puntos de interés específicos, como escuelas, parques o centros comerciales, como variables explicativas en la regresión hedónica (Brasington y Hite, 2005).
	Variables de accesibilidad	Incluyen medidas de accesibilidad, como el tiempo de viaje o la distancia a las estaciones de transporte público, para capturar el efecto del acceso a servicios y empleo en los precios de las viviendas (Anselin y Lozano-Gracia, 2009).
Interacción espacial	Rezagos espaciales	Introducen rezagos espaciales (*spatial lag*) de la variable dependiente y las variables independientes, para abordar la dependencia espacial en los datos (Anselin, 2002).
	Correlación espacial de errores	Consideran la correlación espacial en los términos de error, lo que puede surgir debido a la omisión de variables espaciales no observadas o la presencia de factores espaciales comunes (Dubin, 1998).
Control de la heterogeneidad espacial	Efectos fijos espaciales	Incluyen efectos fijos para áreas geográficas específicas, como barrios o distritos, para controlar la heterogeneidad espacial no observada (Malpezzi *et al.*, 2003).
	Superficie de respuesta espacial	Utilizan funciones de superficie de respuesta, como polinomios espaciales o funciones de base radial, para modelar la variación espacial en las relaciones hedónicas (Paelinck *et al.*, 1979).
Geografía cuantitativa	Regresión geográficamente ponderada (GWR)	Permiten que los coeficientes de las variables explicativas varíen en el espacio, estimando modelos de regresión locales para cada ubicación en el área de estudio (Fotheringham *et al.*, 2003).
	Partición espacial	Dividen el área de estudio en subregiones homogéneas y estiman modelos de regresión hedónica separados para cada subregión (Kim *et al.*, 2003).
	Bayesianos espaciales	Adoptan un enfoque bayesiano para inferir los parámetros de la regresión hedónica, lo que permite incorporar información previa y obtener estimaciones más robustas en presencia de datos escasos o ruidosos (LeSage y Pace, 2009).

Fuente: elaboración propia

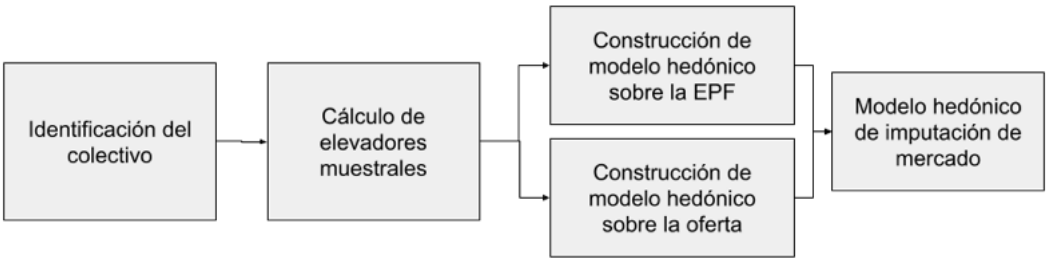
Tabla 2.2. Taxonomía de métodos hedónicos geográficos

2.2.2 Modelo hedónico de mercado

El primer paso en la construcción del modelo de mercado es la correcta caracterización del colectivo de mercado[17], para que sea pueda relacionar con la población de oferta. Primeramente, se definen los distintos estratos que compondrán el colectivo, asegurando que todos tengan suficiente soporte de datos. Dicha estratificación tiene dos dimensiones, una zonal y otra funcional, esta última referida al desglose de características de la vivienda.

Puesto que no es posible relacionar una a una las observaciones de oferta y mercado, al disponer solo de información estadística agregada del mercado, se construye un modelo hedónico de imputación de precios que convierte el precio de oferta de una vivienda al precio de mercado de la misma. Las etapas de las que se compone este proceso se muestran en la Figura 2.5.

Figura 2.5. Etapas del modelo hedónico de mercado



Fuente: elaboración propia.

La necesidad de ajustar el colectivo y realizar una correcta ponderación se justifica por la presencia de sesgos muestrales como la tendencia a sobre-representar a determinados grupos (Särndal *et al.*, 2003), por al desfase temporal entre el momento actual y el recogido en el marco de referencia, o por la falta de respuesta de algún segmento (Lohr, 2019).

Inicialmente se parte de una matriz de diseño o "rejilla"[18] adecuada del colectivo sobre la que se realiza la estratificación (viviendas en régimen de alquiler). Esta segmentación atiende a los criterios zonal y funcional. El zonal, divide la población en diferentes áreas (municipios o barrios), y el funcional, sobre las características de la vivienda (año de construcción, número de habitaciones, etcétera). El diseño cumple, además, dos requisitos:

- La segmentación de la rejilla se construye sobre variables comunes entre

[17] Todas las viviendas que se encuentran en régimen de alquiler.
[18] La rejilla se refiere a la combinación de variables en una matriz de diseño en el proceso de muestreo.

todos los conjuntos de datos (oferta y estadísticas).

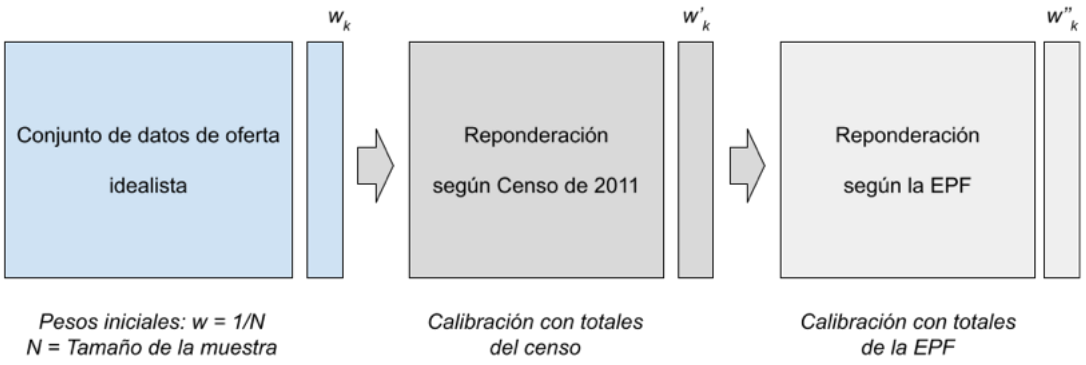
- Cada celda de la rejilla debe disponer de información suficiente, que en nuestro caso se ha establecido en un mínimo de 30 anuncios por celda.

La reponderación se realiza mediante dos procesos de calibración encadenados: el primero, ajusta los elevadores originales de oferta, para adaptarse a la estructura muestral del censo de 2011; el segundo, ajusta los pesos de la primera calibración, para que se adecuen a la estructura poblacional del mercado, recogidos en la EPF. La calibración es un proceso de optimización que busca unos pesos muestrales que se correspondan a la estructura poblacional deseada, pero preservando la estructura original de pesos (Deville y Särndal, 1992).

La calibración requiere trabajar con la misma estratificación en la oferta y en las bases de datos estadísticas, por tanto, se realiza sobre las variables comunes de ambas fuentes. Para el Censo, estos atributos son: número de habitaciones, superficie útil, tipo de municipio, antigüedad, anejos y ascensor.

En el año 2011 la estructura de población del mercado es conocida, al contarse con el dato del Censo de Población y Viviendas. Para los años posteriores (2012 a 2019) se realiza una doble calibración, la primera sobre el censo y la segunda sobre la EPF (véase Figura 2.6). Como los atributos del censo y la EPF no coinciden en todos los mismos, la calibración de la EPF usará: renta media por persona de la zona, tipo de zona, nivel adquisitivo medio de los hogares, tipo de edificio, densidad de población, número de habitaciones y tamaño del municipio.

Figura 2.6. Proceso de cálculo de ponderaciones, doble calibración



Fuente: elaboración propia.

Una vez que se dispone de los pesos, se construye un modelo de conversión que relaciona los precios de oferta con los precios de mercado. Para ello se construyen varios modelos hedónicos que enlazados en cascada: el primero, estima el precio del alquiler según la encuesta de población activa; el segundo, lo hace sobre los

precios en oferta; y el tercero, estima la relación entre ambas magnitudes. Este modelo resultante realiza una estimación del precio de mercado de la vivienda en alquiler medio anual, para cada estrato de la rejilla.

Los modelos de alquiler y de oferta básicos se estiman con el algoritmo *Random Forests*[19] (Breiman, 2001), y cuyos resultados se usan como entrada del modelo de conversión. El primero se calcula sobre los microdatos de la EPF, y el segundo sobre los datos de Idealista. Ambos utilizan las mismas variables independientes y su variable objetivo es el precio anual por metro cuadrado útil. El modelo formulado como una regresión lineal, se indica en la expresión:

$$\begin{aligned} Ln\ \hat{P_m} = {} & \beta_1 \cdot TAMAMU + \beta_2 \cdot TIPOEDIF + \beta_3 \cdot TIPOCASA \\ & + \beta_4 \cdot ZONARES + \beta_5 \cdot SUPERF + \beta_6 \cdot ANNOCON \\ & + \beta_7 \cdot DENSI + \beta_8 \cdot INTERINPSP + \beta_9 \cdot NHABIT \\ & + \beta_{10} \cdot CCAA + \beta_{11} \cdot CAPROV + \beta_{12} \cdot factorGASTOT6 \end{aligned} \quad [2.7]$$

donde P_m representa el precio de mercado de la vivienda[20], $TAMAMU$ el tamaño del municipio, $TIPOEDIF$ el tipo de edificio, $TIPOCASA$ el tipo de vivienda, $ZONARES$ el tipo de zona residencial, $SUPERF$ la superficie útil, $ANNOCON$ el año de construcción, $DENSI$ la densidad de población del área, $INTERINPSP$, $NHABIT$, $CCAA$ la comunidad autónoma[21], $CAPROV$ variable dicotómica que indica si está en la capital, y $factorGASTOT6$ el nivel de gasto del hogar .

Finalmente, se construye un modelo GAM que representa linealmente la correspondencia entre precios de oferta y de mercado. La variable dependiente es el precio de mercado, y usa como covariable el precio de oferta:

$$\begin{aligned} \hat{P_m} = {} & s(\hat{P_o}) + \beta_1 \cdot TAMAMU + \beta_2 \cdot TIPOEDIF + \beta_3 \cdot ZONARES \\ & + s(SUPERF2) + \beta_4 \cdot ANNOCON + \beta_5 \cdot DENSI + \beta_6 \cdot INTERINPSP \\ & + \beta_7 \cdot NHABIT + \beta_8 \cdot CAPROV + \beta_9 \cdot factorGASTOT6 \end{aligned} \quad [2.8]$$

donde los términos son equivalentes a los utilizados en la expresión [2.7], con la salvedad de que los coeficientes de las variables precio de oferta y superficie útil, $s(\hat{P_o})$ y $s(SUPERF)$ respectivamente, se especifican como funciones base de suavizado.

[19] El *Random Forests* es un algoritmo de aprendizaje automático, de la familia de los árboles de regresión, que se explicará con detalle en el capítulo 5.

[20] Para el modelo base de oferta se utilizaría el término $\hat{P_o}$.

[21] Este atributo solo es aplicable al modelo de mercado porque el fichero de microdatos de la EPF sí cuenta con datos de distintas comunidades.

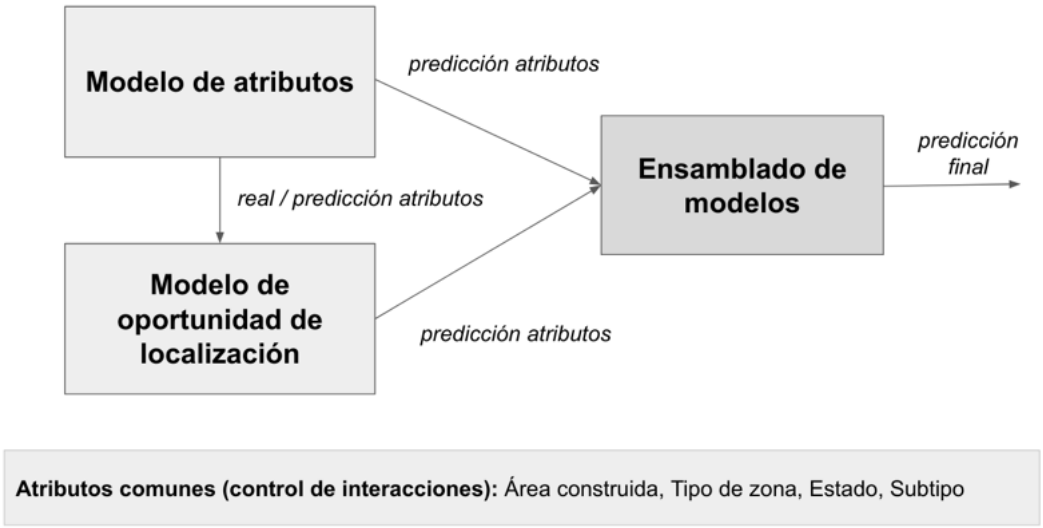
2.2.3 Modelo hedónico de oferta

Dado que las variables utilizadas en el modelo de mercado son limitadas, se construirá un modelo hedónico que recoja las contribuciones de otros atributos (localización, dinámicas de mercado, entre otros). En este caso el precio a predecir será el de oferta.

Se utiliza también el algoritmo *Random Forests*[22], por su mayor capacidad para gestionar las debilidades de los modelos paramétricos (Antipov y Pokryshevskaya, 2012). Para capturar de una forma eficaz las interacciones de los atributos estructurales de la vivienda y los efectos de la localización, se crea un modelo ensamblado que une dos modelos especializados en cada aspecto.

El modelo ensamblado de oferta combina un modelo de atributos y otro de localización. El primero, calcula el precio de la renta mediante un amplio conjunto de características, el de localización toma los errores del modelo de atributos y estima la corrección que se debe realizar sobre el primer modelo dada las características de la zona en la que se encuentra. El resultado es un tercer modelo que ensambla el precio en base a las características y ajusta, en función de la localización, la estimación del precio de mercado. El flujo completo se muestra en la Figura 2.7.

Figura 2.7. Ensamblado de modelos de oferta



Fuente: elaboración propia.

[22] Random Forests es un modelo de aprendizaje automático basado en árboles de decisión o regresión, se describe en detalle en el Anexo 3b del capítulo 3.

Dado que la literatura no es determinante ni en la forma funcional, ni en los atributos a incorporar en el modelado hedónico de la vivienda (Cassel y Mendelsohn, 1985; Freeman, 1979; Rosen, 1974). El proceso se centrará en la cuestión fundamental para lograr un buen ajuste, que como afirma Zyga (2019), es la selección de variables de entrada, por encima del tipo de método de modelización utilizado.

El tipo de variables es muy amplio, aunque se pueden agrupar a cuatro categorías principales (Sirmans *et al.*, 2005): estructurales, de mercado, de localización y de tiempo. En la Tabla 2.3 se describe el detalle de cada una.

Tabla 2.3. Categorías de variables del modelo hedónico de oferta

Categoría	**Motivación**
Estructural	Las características estructurales capturan la contribución de las características físicas de la propiedad, como los metros cuadrados, el número de habitación o el estado de conservación
Características del mercado	Incorpora la principal dinámica de oferta/demanda del mercado donde se ubica el inmueble
Localización	Explicar la contribución de la ubicación en el precio del suelo de un activo, incluye características del vecindario, índices de accesibilidad y otras características geográficas
Dummy de tiempo	Captura el ajuste del precio a lo largo del tiempo, la estacionalidad y los efectos de tendencia

Fuente: elaboración propia

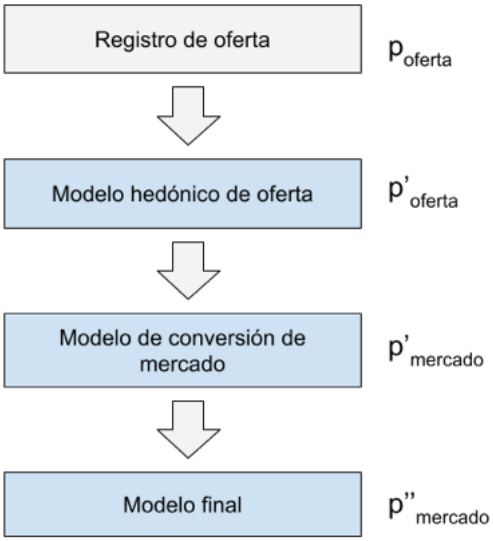
El modelo de localización anterior no utiliza la ubicación geográfica del inmueble, sino que utiliza la utilidad asociada a dicha ubicación. El motivo procede de la justificación del propio modelado hedónico, dado que el precio se explica, parcialmente, por la prima que están dispuestos a pagar los usuarios por la utilidad situacional. Esta utilidad se puede expresar, numéricamente, en términos de accesibilidad[23] a los distintos servicios que tiene alrededor (colegios, hospitales, centros comerciales o centros de trabajo). Si la utilidad situacional es positiva, el demandante de vivienda estaría dispuesto a hacer una puja más alta por la vivienda, si en cambio existen desutilidades, el individuo pujaría por un precio menor (Ottensmann *et al.*, 2008).

[23] Véase para más información acerca de la accesibilidad véase (Batty, 2009).

2.2.4 Modelo hedónico final

El modelo hedónico final calcula el precio de mercado a partir del resultado de los modelos anteriores, como muestra en la Figura 2.8. En un primer paso, se imputa el precio de oferta de los registros originales usando el modelo hedónico de oferta (p'_{oferta}), esto reduce la influencia de las variables omitidas en el uso de los datos originales del portal. Posteriormente, el precio estimado de oferta se transforma en la renta con frecuencia mensual a través del modelo de mercado ($p'_{mercado}$).

Figura 2.8. Etapas del modelado hedónico



Fuente: elaboración propia.

La especificación de este modelo captura la relación funcional entre los precios de oferta y de mercado, pero dado que la fuente no contiene información de zonas concretas, las estimaciones de precios observadas en los estratos zonales muestran algunos comportamientos de precio erráticos[24].

Para corregir el sesgo de la información de las zonas omitidas en el modelo de mercado, el modelo final realiza un ajuste zonal de sesgo de los precios ($p''_{mercado}$), utilizando como fuente de datos auxiliar las series de precios del MITMA (2020).

[24] Principalmente inconsistencia temporal en los precios cuando se desglosan a nivel de municipio o barrio.

2.3 Plano temporal: Índices de precios

Los índices de precios han sido instrumentos de uso común para el análisis económico desde el siglo XIX, pero no fue hasta pasada la primera mitad del siglo XX cuando empezaron a desarrollarse para el análisis financiero.

Los índices de precios de la vivienda (IPV o HPI en inglés) son indicadores clave que ofrecen información sobre el comportamiento de los mercados inmobiliarios. Son una fuente importante de información para los distintos agentes del ecosistema inmobiliario-financiero, lo que los convierte en una herramienta esencial para la toma de decisión de compra de los agentes de mercado (Pollakowski, 1995), y juegan un importante papel en la práctica de políticas macro prudenciales para el control de la formación de burbujas inmobiliarias (Anundsen *et al.*, 2016), dada la relación entre el crecimiento de los precios de la vivienda y el riesgo financiero doméstico.

Su función principal es la de ofrecer mecanismos de información capaces de reducir la incertidumbre y asimetrías de información en el mercado. En la Tabla 2.4 se recogen los distintos beneficios, tanto macro como microeconómicos, derivados de su uso.

Tabla 2.4. Efectos económicos de la utilización de un IPV

Microeconómicos	**Macroeconómicos**
Permite un mejor entendimiento de la influencia de la ubicación, y las dinámicas urbanas	Es un indicador clave de los mercados mercado inmobiliario y de la construcción
Permite a las entidades financieras el potencial de apreciación de los activos inmobiliario.	Permite establecer políticas macro-prudenciales más completas dada la exposición habitual del mercado financiero al sector inmobiliario
Permite estimar mejor el rendimiento futuro de las inversiones en propiedades	Permiten medir el nivel de riqueza de las familias
Ayuda a establecer políticas locales más adecuadas sobre el desarrollo urbanístico, o la mitigación de problemas de accesibilidad de la vivienda	Permiten el control de fenómenos especulativos
Ofrece un mecanismo objetivo para el control de los impuestos urbanos: catastro, transmisiones de bienes inmobiliarios, IBI, entre otros	Es un indicador que sintetiza la percepción de la riqueza de los agentes económicos
Permite conocer el efecto de las políticas de inversión pública que afectan a este tipo de bienes	Es una herramienta para la gestión de la política social asociada a la vivienda, por ejemplo la inversión de interés social

Fuente: elaboración propia

Los construcción de los IPV tiene la dificultad añadida de que, al contrario de

otro tipo de bienes, la vivienda es un activo de carácter heterogéneo, único y cuyas características varían a lo largo del tiempo (renovaciones, deterioros, ampliaciones, etcétera), lo que supone que los precios disponibles para la creación de los índices, generalmente a partir de registros de transacciones[25], son una muestra segada del colectivo de mercado. Para controlar las características a lo largo del tiempo y limitar el efecto de los sesgos, es necesario un proceso de ajuste para hacer que todas las unidades (viviendas) sean comparables entre sí. Si el objetivo de la construcción de índices de precios de vivienda es reflejar el cambio de nivel entre las áreas metropolitanas de una vivienda estándar, Pollakowski (1995) menciona que se sería necesario además, una fuente de datos nacional uniforme con un alto control de calidad. Existe una extensa literatura acerca del ajuste por características o calidad de los productos en los índices, que comienzan con Hofsten *et al*. (1952), Stone (1956), Stigler (1961) y Griliches (1961).

Una cuestión clave a tomar en consideración es que la utilidad de la vivienda tiene un carácter subjetivo y es desconocida para cada individuo, lo que inevitablemente introduce sesgos de omisión de variables. Por ejemplo, el valor percibido de una vivienda podría ser mayor para un individuo que para otro, simplemente por la mayor cercanía a su centro de trabajo o zonas de ocio. Estas preferencias, asociadas a la utilidad, son determinantes para los precios de la vivienda (Ball, 1974; Ball, 1973; Wilkinson, 1974).

En resumen, el proceso de construcción de un IPV debe afrontar dos cuestiones: la primera, un ajuste de calidad y características de los inmuebles; y la segunda, el correcto cálculo de los pesos de los estratos que componen la cesta de viviendas a lo largo del tiempo. En todo caso, los requisitos, rigor y exactitud en la construcción de índices de precios de vivienda dependerán del propósito para el que son construyan.

2.3.1 Tipos de índices de precios

Desde el inicio del siglo XIX se han desarrollado una gran cantidad de índices, los más utilizados son Laspeyres, Paasche, Fisher, cuyos nombres se corresponden con los nombres de sus autores. La definición clásica de número índice tal y como lo define Edgeworth (1925) es: *"[. . .] Un número que a través de sus variaciones indica los aumentos o disminuciones de una magnitud no susceptible de medirse con exactitud [. . .]"*. Es por tanto, un estadístico que mide la variación relativa, en el tiempo o en el espacio, de una magnitud simple o compleja. Dicha magnitud hace referencia en el campo económico a precios, cantidades o valores.

[25] La transacción se refiere a la formalización de una compraventa o de un alquiler.

El formato más sencillo de índice es el denominado como simple, obtenido a través del cociente de los precios entre dos periodos. El denominador contiene el precio en el periodo base, y el numerador el precio para el periodo de estudio (Díaz, 1997). Como el resto de índices de precios, se puede expresar en bases 1 o 100, en el que 1 o 100 se corresponde al valor que toma la magnitud en el periodo base. Su expresión matemática del índice básico I_t para el periodo t sería:

$$I_t = \frac{p_t}{p_0} \quad [2.9]$$

Donde p_t es el precio en un periodo dado t, y p_0 el precio en el periodo base. Dado que las magnitudes a medir en general son complejas y un sólo índice debe sintetizar el precio de un elemento altamente heterogéneo, es necesario establecer dos criterios: el de agregación y el de ponderación.

El criterio de agregación atiende a determinar la forma en la que se sintetizan las variables en una sola magnitud (como el poder adquisitivo o nivel general de precios). Las variables son los precios de los distintos elementos que componen una cesta de productos, por ejemplo los precios de mercado de distintos productos de gran consumo en el IPC. La agregación se realiza a través de promediar estos valores, por ejemplo, con una media aritmética o una geométrica ponderada.

Por otro lado, el criterio de ponderación consiste en atribuir un determinado peso a cada una de las variables que se promedian, es decir, dándoles la importancia relativa con respecto al grupo al que pertenecen. La opción más simple sería dar el mismo peso a todos los elementos que componen el índice, otra más elaborada sería hacerlo en función del peso poblacional de los estratos que componen la muestra.

Si se combinan todos los posibles criterios de agregación y ponderación, se podrían obtener un conjunto muy numeroso índices. De hecho, Fisher (1922a) hace referencia a 134 fórmulas distintas para calcular números índices. En general las más comunes son las cinco siguientes:

- Índice de Laspeyres (I_{Lo}): es una media aritmética de indices de precios simples cuyas ponderaciones son el valor de las transacciones p_{it} o cantidades realizadas en el período base q_{i0}. Es posiblemente el índice más usado por las agencias nacionales de estadística, debido a las dificultades para obtener datos sobre cantidades o gastos del período actual. Su método de calculo es:

$$I_L = \frac{\sum_{i=1}^{n} p_{it}\ q_{i0}}{\sum_{i=1}^{n} p_{i0}\ q_{i0}} \quad [2.10]$$

- Índice de Paasche (I_P): es también una media aritmética de índices simples,

aunque usa como coeficiente de ponderación el precio de las transacciones efectuadas en el período actual calculado a precios del período base, p_{i0}. El índice de Paasche es una media agregativa de precios ponderados por las cantidades del período actual:

$$I_P = \frac{\sum_{i=1}^{n} p_{it}\ q_{it}}{\sum_{i=1}^{n} p_{i0}\ q_{it}} \qquad [2.11]$$

- Índice de Lowe (I_L): de uso común en muchas agencias de estadística, se obtiene al definir el índice como el cambio porcentual en el coste total de adquirir una "cesta de productos" entre los periodos comparados. En este caso las cantidades de cada estrato i se fijan de antemano y no proceden de ningún periodo (q_i). Este tipo de índice fue propuesto por primera vez por Lowe (1824), y se calcula según la siguiente expresión:

$$I_{Lo} = \frac{\sum_{i=1}^{n} p_{it}\ q_i}{\sum_{i=1}^{n} p_{i0}\ q_i} \qquad [2.12]$$

- Índice de Marshall-Edgeworth: para mitigar la debilidad de los índices de precios más habitual en Laspeyres y Paasche, que es la sobreponderación de precios o cantidades, Marshall y Edgeworth (1888) propusieron un índice que pondera los precios por la media aritmética de las cantidades, definido como:

$$P_{ME} = \frac{\sum[p_{c,t_n} \cdot \frac{1}{2} \cdot (q_{c,t_0} + q_{c,t_n})]}{\sum[p_{c,t_0} \cdot \frac{1}{2} \cdot (q_{c,t_0} + q_{c,t_n})]} \qquad [2.13]$$

- Índice de Fisher: Fisher (1922a) define el denominado índice ideal, calculado como media geométrica de los índices de Laspeyres y Paasche:

$$I_F = \sqrt{I_L \cdot\ I_P} \qquad [2.14]$$

Los índices de Laspeyres y Paasche, según se definen en las expresiones [2.10] [2.11], se pueden entender como razones de valores agregados. El índice de precios de Laspeyres mide cuánto cuesta adquirir la cesta de bienes $(q_{t0}...q_{n0})$, mientras que el índice de Paasche mide el valor de la cesta $(q_{t0}...q_{nt})$.

Todos los índices anteriores se calculan sobre un periodo base, denotado como t_0, que puede ser fijo o móvil. Cuando el periodo base es dinámico, estos índices se denominan encadenados. Por ejemplo, podemos definir el índice de Laspeyres para un periodo t como el producto de una sucesión de índices de Laspeyres encadenados:

$$P_{tn} = \frac{\sum_{i=1}^{n} p_{i,t_1}\ q_{i,t_0}}{\sum_{i=1}^{n} p_{i,t_0}\ q_{i,t_0}} \times \frac{\sum_{i=1}^{n} p_{i,t_2}\ q_{i,t_1}}{\sum_{i=1}^{n} p_{i,t_1}\ q_{i,t_1}} \times \cdots \times \frac{\sum_{i=1}^{n} p_{i,t_n}\ q_{i,t_{n-1}}}{\sum_{i=1}^{n} p_{i,t_{n-1}}\ q_{i,t_{n-1}}} \quad [2.15]$$

Existen cuatro aproximaciones teóricas de índices de precios y cantidades: axiomática, económica, estocástica y estadística (Balk, 2008). La axiomática, establece un conjunto de propiedades deseables para un índice (Diewert, 1976). En el enfoque estadístico, los índices se construyen mediante conceptos estadísticos, como la regresión y la descomposición de varianzas (Court, 1939). La estocástica (Theil, 1967), considera que los cambios en los precios y las cantidades son resultado de procesos aleatorios, y se basa en la teoría de la probabilidad para construir los índices. Finalmente, en la económica el índice representa el comportamiento de los agentes económicos (Konüs, 1924), y permite vincularlo a conceptos económicos fundamentales. Estos índices se derivan de funciones de utilidad o de costes que representan las preferencias o tecnologías de los agentes. Por ejemplo, un índice de precios de Laspeyres puede interpretarse como el costo mínimo para alcanzar un nivel de utilidad fijo cuando los precios cambian.

Desde el ángulo axiomático el índice debería cumplir una serie de propiedades (véase el Anexo 2a del presente capítulo). Los índices que cumplen el mayor número de estas propiedades se denominan índices superlativos. Entre los cinco tipos mencionados anteriormente, el de Fisher es el único que podría considerarse como tal (Auer y Wengenroth, 2020; Diewert, 1976; Triplett, 1996).

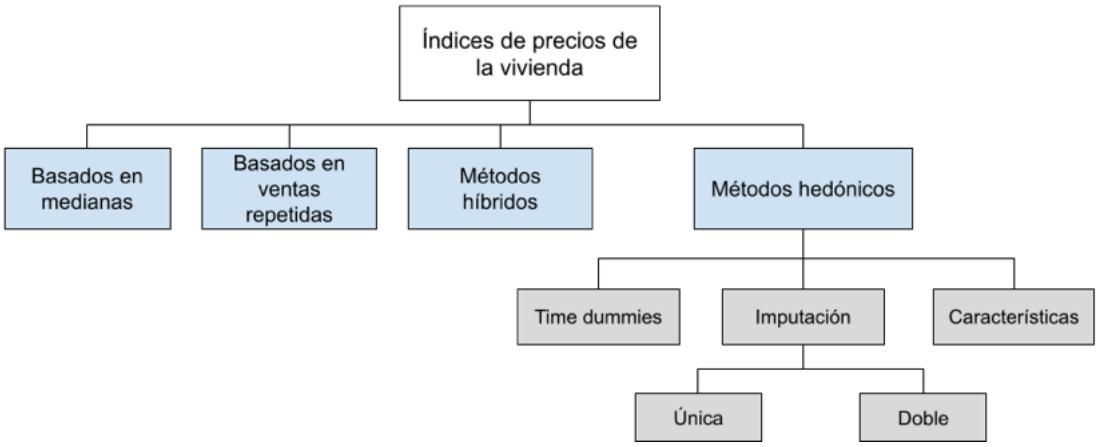
Los métodos mencionados anteriormente se basan en ratios y productos entre precios y cantidades, muchos de ellos están recogidos en el trabajo de Fisher (1922a). No obstante, existen alternativas como los índices basados en diferencias (Diewert, 2005). Esta línea iniciada por Bennet (1920) y Montgomery (1929), descompone una diferencia de valor en la suma de una diferencia de precio más una diferencia de cantidad.

Hay al menos cinco áreas generales en las que son de aplicación este tipo de índices, cuatro de ellas aplicadas a las finanzas empresariales y una quinta a hogares: descomposición de los cambios en los ingresos, descomposición en de los cambios en costes, descomposición de los cambios en beneficios, análisis de varianza y por último los cambios en el superávit de los consumidores (Diewert *et al.*, 2020).

2.3.2 Métodos de construcción de índices de precio de la vivienda

En su revisión de las distintas formas de construcción de un IPV, Hill (2013) identifica siete métodos, entre los que se destacan los cuatro principales mostrados en la Figura 2.9, que son: basados en mediadas, basados en ventas repetidas, métodos híbridos y métodos hedónicos. En la familia de los hedónicos, las tres aproximaciones que se plantean son: las basadas en "*dummies*"[26] de tiempo, las basadas en imputación y las de características.

Figura 2.9. Taxonomía de métodos de índices de precio de la vivienda



Fuente: elaboración propia a partir de Hill (2013).

Ninguno de los métodos anteriores se ha impuesto al resto, quizá porque la disponibilidad de fuentes y capacidad de proceso de la información condiciona qué métodos se pueden aplicar (Eurostat, 2014). Por ejemplo, Finlandia y España, utilizan métodos hedónicos para la construcción de los índices. Esto es debido a que hay numerosas fuentes con un importante desglose que lo permiten. En cambio, el método de ventas repetidas es casi exclusivo de los Estados Unidos, por cuestiones ligadas a su mercado de la vivienda. Se trata de un mercado muy dinámico, en la que existe mucha mayor movilidad de personas si la comparamos con Europa. En Latinoamérica, en cambio hay un menor número de índices sofisticados que en países de la Unión Europea.

2.3.2.1 Modelos basados en medianas

Es el tipo más simple y se desarrolla como un número índice sobre los precios medianos. Este índice se calcula según la siguiente expresión analítica:

[26] Una variable *dummy* es una variable ficticia dicotómica.

$$I_t = \frac{med(P_t)}{med(P_0)} \quad [2.16]$$

donde $med(P_t^e)$ es la mediana del precio de la vivienda para un estrato de la muestra e, en el momento del tiempo t. Mientras que $med(P_0^e)$ es la mediana para dicho estrato en el periodo base.

Los principales atractivos de los índices medianos son su menor necesidad de datos, su mayor simplicidad de cálculo y su mejor facilidad de comprensión. Su principal desventaja es que están sujetos a sesgos de distinta naturaleza, como que pueden confundir cambios en los precios con diferencias de calidad y, por lo tanto, pueden proporcionar estimaciones temporalmente inestables en cuanto a cambios del precio de la vivienda. Por ejemplo, los cambios de composición de la muestra de un periodo pueden introducir variaciones incontroladas en los precios medianos, que no son atribuibles a un cambio fundamental en el mercado sino a la aleatoriedad de que existan más inmuebles de un tipo que de otro.

Existen una gran cantidad ejemplos como son los índices de Mediana Metropolitana de la Asociación Nacional de Agentes Inmobiliarios (NAR) en los Estados Unidos, los índices del Instituto de Bienes Inmuebles de Australia (REIA) y *LJ Hooker / BIS Shrapnel* en Australia (Hill *et al.*, 2018). En el ámbito de los precios de oferta en España encontramos, por ejemplo, los índices de precios de oferta de los portales inmobiliarios Fotocasa e Idealista.

Para reducir el efecto de cambios de composición existe una versión más elaborada del índice de medianas que permite controlar estas variaciones. Esta modalidad, denominada de ajuste mixto, estima el precio del estrato como una media ponderada de medianas. Cada una ellas se corresponde a un subestrato muestral, que representa diferentes grupos de características o niveles de calidad de los inmuebles.

2.3.2.2 Método de ventas repetidas ponderadas

El primer método desarrollado para la elaboración de índices de precios de vivienda fue el de ventas repetidas (*repeat sales*) desarrollado por Bailey (1963), aunque Shiller (2008) atribuye el origen del método a Wyngarden (1927) y Wenzlick (1952). Este índice, rescatado por Case (1986), y ampliado casi en la versión que conocemos por Case y Shiller (1987) en su artículo *Prices of Single – Family Homes since 1970: New Indexes for four Cities*, fue posteriormente adaptado y modificado por la OFHEO, y es hoy en día de hoy el método aplicado en el índice de referencia inmobiliario S&P/Case Shiller, en los Estados Unidos.

La ventaja de este método radica en el hecho de que al utilizar información de

precios de las mismas viviendas en dos puntos del tiempo, se pueden controlar mejor las diferencias entre los atributos de las distintas propiedades, sin tener que estimar directamente sus contribuciones marginales.

Este método se compone de un procedimiento de regresión en tres etapas. La primera, construye una regresión simple entre el logaritmo del cambio relativo en los precios observados entre la segunda y la primera transacción, que actúa como variable explicada, frente a un conjunto de variables *dummy*, una por cada periodo de tiempo de la muestra.

La variable *dummy* D toma para cada vivienda el valor cero en todos los periodos, excepto en los periodos en se producen las dos ventas. Para el periodo de la primera venta, la *dummy* toma el valor de -1, y para el periodo de la segunda, la *dummy* toma un valor de 1. La especificación funcional de esta primera etapa vendría especificada como:

$$\Delta P_i = \sum_{t=0}^{T} \beta_t D_{i,t} + \epsilon_i \qquad [2.17]$$

donde ΔP_i es el logaritmo del cambio relativo de precios observados P entre la primera y segunda transacción para una vivienda i. $D_{i,t}$ se corresponde a la variable *dummy* para la vivienda i en el periodo t y ϵ los residuos aleatorios del modelo.

Una vez calculada la regresión, se toma el vector de residuos al cuadrado (ϵ^2), sobre los que se realiza una regresión ponderada, para el tiempo transcurrido $(t-s)$ entre la primera y segunda transacción. El término constante de la regresión es un estimador de $2 \cdot \sigma_N^2$ (dos veces la varianza del error aleatorio para cada vivienda).

En esta segunda etapa, el coeficiente de la pendiente se estima a partir de la varianza del cambio trimestral (en el término de camino aleatorio gaussiano *C*), y cuya estimación vendría definida por:

$$E[\epsilon] = A \cdot (t-s) + B \cdot (t-s)^2 + 2 \cdot C \qquad [2.18]$$

La tercera etapa, es una regresión de mínimos cuadrados generalizados calculada de forma iterativa sobre la primera regresión, después de ponderar para cada observación por la raíz cuadrada del valor ajustado en la segunda etapa. La expresión correspondiente sería:

$$\frac{\Delta P}{\sqrt{\hat{d}_i^2}} = \sum_{\tau=0}^{T} \beta_\tau \frac{D_{i\tau}}{\sqrt{\hat{d}_i^2}} + \frac{\epsilon_i}{\sqrt{\hat{d}_i^2}} \qquad [2.19]$$

Finalmente, el índice se construye como:

$$I_t = 100 \cdot e^{\hat{\beta}_t} \quad [2.20]$$

donde $\hat{\beta}_t, t = 1, 2, 3...T$ son los parámetros estimados por mínimos cuadrados generalizados. Es recomendable un ajuste por log-normalidad de la distribución, dando lugar al siguiente índice ajustado:

$$I_t = 100 \cdot e^{\hat{\beta}_t + 0.5\ \sigma^2_{\hat{\beta}}} \quad [2.21]$$

A diferencia de los índices basados en medianas, esta metodología arroja un error estándar asociado a cada estimación $\sigma_{I_t} = I_t \cdot \sigma_{\hat{\beta}_t,}$. Aunque este error puede representar ruido, se ha demostrado que, al ser consistente en el tiempo, no se afecta a la estimación del índice.

En este método, las series necesarias para implementar un índice de precios de vivienda deben contener, al menos, la ubicación exacta, el tipo de vivienda, el precio y la fecha de cada una de las transacciones. Para el intervalo de tiempo utilizado, debe existir una muestra suficientemente grande que garantice un número significativo de ventas repetidas y uniformes.

Esta aproximación cuenta con algunos inconvenientes, principalmente por la indisponibilidad de información y por cambios de calidad en las viviendas entre los momentos de primera y segunda venta. El primer problema se refiere a que no todas las viviendas están disponibles en el mercado para cualquier periodo t de cálculo. El segundo, que si se producen cambios de estructura en la vivienda, obligan su exclusión del cálculo por no corresponder al mismo bien observado en el periodo t_k.

Surge entonces el reto de controlar los cambios de composición a lo largo del tiempo y del espacio, por ello es necesario utilizar un método que permita controlar por calidad los atributos de dichos bienes. Una forma de resolverlo es tomar exclusivamente las variaciones de los precios de los productos que no han sufrido variaciones cualitativas significativas en el periodo considerado (Shiller, 1991). Aunque esto plantea el inconveniente de generar problemas en la estabilidad del índice, debido a que las restricciones en el tamaño muestral, especialmente en bienes tan heterogéneos como la vivienda, puede dar lugar a sesgos de composición importantes.

2.3.2.3 Índice de precios hedónicos

Otra aproximación para la construcción de índices de precios de vivienda sería a través del uso de métodos de precios hedónicos. Este método se basa en los trabajos de Griliches (1961), Rosen (1974), Berndt (1991), y Berndt y Rappaport (2001), y es de especial interés para la vivienda por su naturaleza de bien singular. Otra de las ventajas del método hedónico es que permite controlar la no respuesta, ya que el precio de los elementos de la cesta de viviendas se estima por un modelo hedónico.

El número índice se construye con los precios estimados del modelo, como se observa en el siguiente ejemplo: se construye un índice hedónico de Laspeyres I_t para un momento del tiempo t, cuyas contribuciones proceden de un modelo hedónico lineal con forma semilogarítmica:

$$I_t = \frac{e^{\alpha_t + \sum_{J=1}^{n} \beta_{j,t} \cdot \overline{log(X_{j0})}}}{e^{\alpha_0 + \sum_{J=1}^{n} \beta_{j,0} \cdot \overline{log(X_{j0})}}} \qquad [2.22]$$

donde $\overline{log(X_{j0})}$ es el valor medio de la característica j en el año base 0 (representando en realidad las cantidades en el periodo base $t = 0$).

Dentro de los métodos de índices hedónicos existen tres modalidades, la basada en "*dummies*" de tiempos (Court, 1939), la basada en características y la de imputación.

El método de *dummies de tiempos* ha sido utilizado ampliamente en la academia, pero escasamente por organismos oficiales (Eurostat, 2014). Captura los efectos fijos temporales a través de variables de tipo *dummy*, cuyos coeficientes de regresión recogen la variación de los precios atribuida al tiempo, la forma funcional se define según la expresión:

$$log(P) = \beta_0 + \sum_{t=1}^{T} \delta_t \; D_t + \sum_{t=1}^{T} \beta_k \; Z_k + \varepsilon \qquad [2.23]$$

donde P representa el precio, D_t las variable *dummy* de tiempo para el momento t y Z_k el valor de la característica k. Por tanto la exponencial del coeficiente de la variable *dummy*, $P_{0,t}^{TD} = exp(\hat{\delta}_t)$, ofrece una medida del efecto del cambio temporal desde el momento 0 al t. Este enfoque tiene el inconveniente de que a medida que se incorporan nuevos periodos ($t + 1$) se deben re-estimar los coeficientes calculados de los periodos anteriores ($1..t$).

El método de imputación resuelve el problema de la ausencia de información para ciertos estratos a lo largo del tiempo, puesto que es común que la muestra no cuente con información completa para todos los estratos, y en todos los periodos.

En el caso de un índice de Laspeyres se imputan los precios de la cesta de viviendas del periodo base con los precios del periodo t, estimados con el modelo hedónico:

$$I_L = \frac{\sum_{i=1}^{n} \hat{p}_{i,t}\ q_{i,0}}{\sum_{i=1}^{n} p_{i,0}\ q_{i,0}} \quad [2.24]$$

donde $\hat{p}_{i,t}$ es el precio estimado para el periodo t y el estrato i, $\hat{p}_{i,0}$ el precio observado para el periodo 0 y estrato i, y $\hat{q}_{i,0}$ la cantidad del estrato i para periodo 0. Este primer método, que se denomina de imputación simple, tiene una variante en la que el precio del momento base también se imputa a través del modelo. La expresión sería por tanto:

$$I_L = \frac{\sum_{i=1}^{n} \hat{p}_{i,t}\ q_{i,0}}{\sum_{i=1}^{n} \hat{p}_{i,0}\ q_{i,0}} \quad [2.25]$$

donde $\hat{p}_{0,t}$ sería el precio estimado para el periodo base y estrato i.

Varios autores, como Hill (2013), recomiendan la imputación doble porque evita sesgos de omisión de variables.

Existe un método adicional, denominado de características (Eurostat, 2014; Hill, 2013), que se podría considerar como un caso particular de imputación doble. Utiliza un modelo hedónico para calcular el precio medio de cada estrato, definido por una combinación de características (Eurostat, 2014).

2.3.2.4 Método de precios híbridos

El modelo híbrido fue sugerido por Case y Quigley (1991) y desarrollado con más profundidad por Quigley (1995). La idea central consiste en combinar la estimación por el método de ventas repetidas y el método de precios hedónicos, para lograr un mayor control en las valoraciones. Este control es necesario ante los posibles cambios de composición en la cesta de inmuebles utilizada, o por cambios en la estructura cualitativa de las viviendas (por cambios tecnológicos, constructivos, etcétera).

Este modelo estima el precio con dos expresiones sobre las que se aplican una serie de restricciones a los parámetros comunes. La siguiente expresión representa el componente hedónico, y se estima sobre todas las transacciones de propiedades residenciales que se vendieron una vez, durante el periodo de estudio.

$$log(Y_t) = log(A) + \beta_1 log(X_{1t}) + \beta_2 log(X_{2t}) + \sum_{n=1}^{t} \gamma_n + \epsilon \quad [2.26]$$

donde Y_t es el precio de una propiedad vendida una vez durante el periodo de estudio en el periodo t, o el precio en el momento de la segunda transacción en cualquier par de transacciones consecutivas sobre una propiedad; el termino A es la intersección; τ es el precio en el momento de la primera transacción en los pares de operaciones consecutivas; X_{1t} y $X_{1\tau}$ son características continuas de la propiedad (como la superficie total construida o el tamaño de la parcela); y X_{2t} y $X_{2\tau}$ son las características discretas (como el número de baños, habitaciones) en el momento de la transacción.

La siguiente expresión refleja el componente de ventas repetidas sobre el cambio en los atributos, y se estima para todos los pares de transacciones consecutivas que hayan entrado al mercado más de una vez, durante el periodo de estudio.

$$log(Y_t) = log(Y_\tau) + \beta_1\ log(\frac{X_{1t}}{X_{1\tau}}) + \beta_2(X_{2t} - X_{2\tau}) + \sum_{n=1}^{t} \gamma_n + \epsilon \qquad [2.27]$$

Una vez calculadas las ecuaciones, el índice de precios se calcularía a partir del vector de coeficientes compuesto de estimaciones en el cambio del índice de precios $\gamma_n, n = 1, ..., T$, para cada periodo n.

2.3.2.5 Índices basados en fuentes de datos alternativos

El uso de datos de Internet para construir índices alternativos ha empezado a tomar cada vez más importancia. Quizá ha sido la crisis del Covid-19 la que ha disparado la necesidad de incorporar nuevas fuentes de datos. Diversos centros de estudios y agencias estadísticas nacionales han comenzado a replantearse el uso exclusivo de fuentes tradicionales (Biancotti *et al.*, 2020).

Aunque el uso de fuentes no oficiales para el cálculo de índices de la vivienda no es algo nuevo, una de las primeras referencias de índice basado en fuentes alternativas es el índice de compraventa y alquiler calculado sobre una base de datos de anuncios clasificados en el periódico "El Mercurio", construido en Santiago de Chile con una serie de datos mensual desde 1998 a 2002 (Desormeaux y Piguillem, 2003). Más cercano en el tiempo, Anenberg y Laufen (2017) desarrollan un índice de precios altamente actualizado para la Reserva Federal de Estados Unidos, sobre datos de oferta de múltiples asociaciones de agencias inmobiliaras (MLS[27]) como en transacciones inmobiliarias. La ventaja de utilizar esta fuente, según sus los autores, es que por un lado recoge las condiciones actualizadas y detalladas del mercado y, por otro, adelanta comportamientos de los índices oficiales, en este caso el índice anticipaba el

[27] MLS, acrónimo de "Multiple Listing Services", se refiere a las asociaciones entre agencias inmobiliarias estadounidenses.

comportamiento del índice Case-Shiller[28], con varios meses de antelación .

Los retrasos de meses en índices oficiales limitan la capacidad de reacción ante situaciones de choque en el mercado. Este retraso también genera un desequilibrio de información entre lo que se conoce del mercado y la situación real. Por ejemplo, en Estados Unidos, la publicación de los datos de los índices de precio al consumo tienen un efecto inmediato en los mercados de valores de las compañías cotizadas, a pesar de que esta información proceda da una situación de meses pasados (Anenberg y Laufer, 2017).

Incluso bancos centrales, como es el caso de Italia, han desarrollado análisis sobre el potencial de esta información. Loberto (2018) utiliza información del portal Inmobiliare, con frecuencia semanal, para construir un índice de la vivienda alternativo. En este estudio, se observa que los índices de oferta tienen una alta correlación con los índices basados en transacciones (un R^2 de 0,96 con respecto a los datos registrados oficialmente[29]). También evidencian problemas al trabajar con este tipo de información, como que una misma propiedad pueda estar anunciada más de una vez, o la ausencia de valores en ciertos campos.

En el artículo del Banco de Francia, Bricongne, Meunier y Syvain (2023) realizan un seguimiento de los precios con frecuencia diaria sobre datos de cinco portales en el Reino Unido. Mediante técnicas de aprendizaje automático desarrollan un modelo de correspondencia entre precios de oferta y los registrados por los notarios. Del mismo modo, en Asia, Wang, Li y Wu (2020) elaboran un índice de precios de la vivienda para 274 ciudades en China basándose en datos de portales inmobiliarios, y Clark (2018) desarrolla un modelo sobre datos del portal Zoopla, en Inglaterra.

Pero no solamente existen referencias basadas en datos alternativos para construir índices de precios. Chauvet *et al.* (2013) y Alexander *et al.* (2014) construyeron índices basados en las búsquedas más habituales en Google Trends[30] sobre el mercado inmobiliario y su regulación. En él demuestra como de un índice de “sentimiento” en Internet puede estar altamente correlacionado con la evolución de los precios. Asimismo, Galesi *et al.* (2020) usa datos de portales inmobiliarios para estudiar la relación de precios de oferta y de transacción, y estimar el poder de negociación de los agentes.

[28] El índice Case-Shiller es un índice de precios mensual de la vivienda de referencia en los Estados Unidos de América, compuesto por los precios de vivienda de las diez principales áreas metropolitanas del país.
[29] Datos estadísticos basados en registros oficiales (OMI) del Ministerio de Hacienda Italiano.
[30] Google Trends es una herramienta muestra los términos más frecuentes en su motor de búsqueda.

2.3.3 Índices de precio de la vivienda en España

Las fuentes de información de precios de la vivienda disponibles en España proceden tanto de la administración como de entidades privadas. La primera, comenzó a publicar precios de compraventa en la década de 1980, y a partir 2019, inició la distribución de las primeras estadísticas detalladas para el alquiler. Las principales entidades que ofrecen series de precios de la vivienda son: el Banco de España, el INE, el Ministerio de Transportes, Movilidad y Agenda Urbana (anteriormente Ministerio de Fomento). y otros organismos privados, como empresas de tasación y portales inmobiliarios.

Las primeras estadísticas de precios de la vivienda de compraventa fueron publicadas por el Ministerio de Fomento en 1987. Posteriormente, en 2003, el Ministerio de Vivienda asumió esta responsabilidad, para que en el año 2006, el INE (2006b), bajo el encargo de Eurostat, creara una metodología de índices de precio estandarizada que formaría parte del IPC (López, 2007). En 2009, el INE desarrolló el primer índice de la vivienda (IPV) que utilizaba datos de las transacciones de compraventa registradas por los notarios.

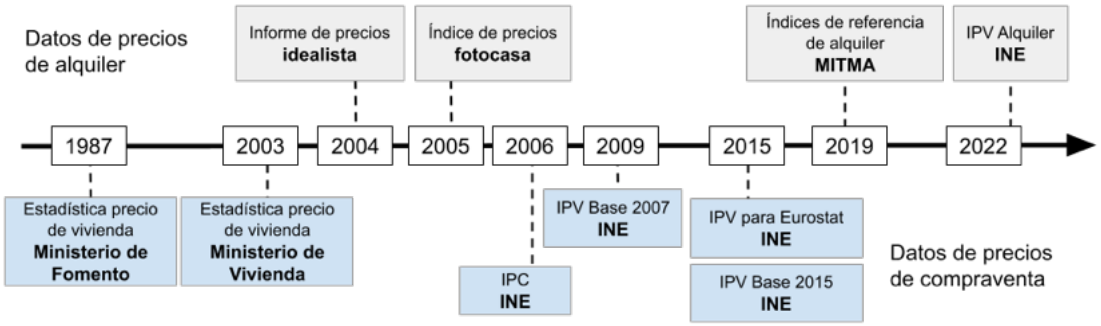
Posteriormente, en 2015, el INE actualizó la metodología para integrarse en el índice de precios europeo armonizado de la vivienda de Eurostat, momento en el que actualizó la metodología anterior del IPV. En 2019, el Ministerio de Transportes, Movilidad y Agenda Urbana publicó la primera base de datos del alquiler desglosada de ámbito nacional[31]. Posteriormente, en 2021, el INE publicó el Índice del Precio de la Vivienda en Alquiler (IPVA) en su espacio de estadísticas experimentales.

Desde el lado privado, los dos portales inmobiliarios más utilizados comenzaron a publicar estadísticas de los precios de las viviendas publicados en sus plataformas. Idealista comenzó a publicar datos de compra y alquiler en 2004 y Fotocasa lo hizo a partir de 2005.

La evolución histórica mencionada anteriormente se resume en la Figura 2.10, en dicha figura se incluye la incorporación de datos de los mercado inmobiliarios de compraventa y de alquiler.

[31] Excluyendo las tres provincias del País Vasco y Navarra.

Figura 2.10. Evolución histórica de la publicación datos de precios de la vivienda en España



Fuente: elaboración propia.

Las seis fuentes principales de precios de la vivienda residencial se resumen en la Tabla 2.5. Cuatro de ellas contienen información de alquiler y dos son exclusivamente de compraventa. Solo las series construidas por el INE se pueden considerar índices de precios desde un punto de vista estrictamente metodológico, el resto son estadísticas de precios con distinto nivel de desglose. Los tres índices son de tipo Laspeyres encadenado, y para el caso de compraventa, utilizan ajuste hedónico.

Tabla 2.5. Resumen fuentes de datos con precios de la vivienda en España

Nombre	Entidad	Descripción	Mercados	Método / Ajuste
Índice de Precios del Alquiler	MITMA	Medianas de los precios de alquiler declarados en IRPF y Catastro	alquiler	Medianas / Ninguno
IPV	INE	Índice del precio de la vivienda de compraventa, calculado sobre las transacciones registradas por el colegio de notarios	compraventa	Laspeyres / Hedónico
IPV Armonizado	Eurostat / INE	Índice del precio de la vivienda de compraventa armonizado según Eurostat	compraventa	Laspeyres / Hedónico
IPV Alquiler	INE	Índice de precio de la vivienda en alquiler de tipo experimental	alquiler	Laspeyres
Informe de Precios	idealista	Informe de precios de la oferta en idealista	compraventa y alquiler	Medianas / Mixto
Índice Inmobiliario	Fotocasa	Informe de precios de la oferta en Fotocasa	compraventa y alquiler	Medias / Ninguno

Fuente: elaboración propia

Fuente: elaboración propia

2.3.3.1 Índices publicados por el Banco de España

El Banco de España, ofrece una visión completa del mercado a través de los Indicadores del Mercado Inmobiliario (2022b), como parte del conjunto de indicadores económico-financieros que ofrece mensualmente. Las series de datos, puramente inmobiliarias, incluyen información de precios, volumen de inmuebles en oferta, demanda, condiciones de financiación y datos de accesibilidad financiera de los hogares para la compra de vivienda.

La información de precios procede de diversas fuentes y su frecuencia depende de cada uno de los indicadores (en algunos casos redistribuyen el dato de otras entidades como el MITMA o el INE). Los datos publicados son: el índice de precios de la vivienda, trimestral (elaborado por el INE); el valor tasado de la vivienda libre por metro cuadrado (recopilado por el MITMA); los precios de oferta mensuales de los portales Idealista, fotocasa; datos semestrales de la tasadora Sociedad de Tasación; el IPC de alquileres mensual (INE); el índice de costes mensual de la edificación del MITMA; y el deflactor trimestral de la inversión en la vivienda, a partir de la Contabilidad Nacional (elaborado por el INE).

Los datos de oferta incluyen solo viviendas de obra nueva puestas en mercado, los de demanda, a las operaciones de compraventa (registradas por el Registro de la Propiedad) y el volumen nacional de inversión en vivienda. Por otra parte, los datos de accesibilidad ofrecen información sobre plazos y tipos de interés de las hipotecas, las tasas de esfuerzo de compra de vivienda y la riqueza inmobiliaria de los hogares.

2.3.3.2 Índices publicados por Ministerio de Transportes, Movilidad y Agenda Urbana

El ministerio encargado de la vivienda ha sido tradicionalmente el de Fomento, aunque recientemente ha cambiado de nombre para denominarse Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA).

El MITMA ofrece, de forma periódica, información estadística del mercado inmobiliario y constructor, y en paralelo gestiona dos observatorios relacionados: el de la vivienda y del suelo, que publica un boletín trimestral analizando la evolución del mercado inmobiliario y la construcción residencial; y el de la vulnerabilidad urbana, que estudia y controla los fenómenos de inseguridad de los ciudadanos para el acceso a la vivienda.

Desde un punto de vista estadístico ofrece información de la construcción: licencias, costes y percepción coyuntural del sector, además de series trimestrales de precios de tasación de viviendas en compraventa y precios del suelo urbano.

En 2019 comenzó a publicar el "Sistema Estatal de Índices de Referencia del Precio del Alquiler de Vivienda", que contiene medidas estadísticas de los precios la vivienda en alquiler en España. Responde a la inexistencia de fuentes oficiales de precios con fines regulatorios y de control, que tal y como define el MITMA, persigue a tres metas: (1) garantizar la transparencia; (2) servir a la aplicación de políticas públicas que fomenten la oferta de vivienda asequible, y (3) facilitar la planificación de la agenda urbana (MITMA, 2020). Fue desarrollado un grupo técnico coordinado por el Ministerio de Fomento, en el que participaron: la Agencia Tributaria, la Dirección General del Catastro, el Instituto Nacional de Estadística, el Banco de España, el Colegio de Registradores y el Departamento de Asuntos Económicos del Ministerio de Presidencia.

La base de datos previa no abarca todo el territorio de España, ya que excluye información de las provincias de Guipúzcoa, Vizcaya, Álava y Navarra. En cuanto al desglose geográfico, la información se encuentra disponible en cinco niveles: sección censal, distrito censal, municipio, provincia y comunidad autónoma. Además, en el aspecto temporal, los datos se analizan anualmente, desde el año 2015 hasta el 2020.

Desde el punto de vista técnico, la explotación estadística combina entre fuentes tributarias y catastrales. Requiere que la referencia catastral suministrada sea válida, así como los datos del bien inmueble y su titularidad. Además, revisa que las rentas generadas cumplan determinadas validaciones, para evitar registros erróneos o atípicos. La medida que se utiliza son la "Renta", definida como el *"..resultado de dividir los ingresos íntegros anualizados declarados en el IRPF por la superficie en m² de la vivienda (expresada en Euros/m² mes).."*, y la "Cuantía" que son los ingresos íntegros anualizados (MITMA, 2020).

La muestra está formada por todos los bienes inmuebles que incluyen algún local de uso vivienda, que se hayan registrado ingresos por arrendamiento como vivienda habitual en el modelo 100 del IRPF[32].

Las variables disponibles en la estadística son que ofrece el fichero son:

- Número de bienes inmuebles arrendados para vivienda habitual
- Mediana del precio por unidad de superficie en € / m² y total.
- Percentiles 25 y 75 del precio por unidad de superficie en € / m² y total.

Esta fuente de información es de gran utilidad, aunque cuenta con algunos inconvenientes: el primero, que las declaraciones de la cuantía del alquiler se hacen por ejercicio fiscal, por lo que una renta que solamente está en vigor durante 3 meses es indistinguible de una renta que está activa durante el año

[32]Impuesto sobre la Renta de las Personas Físicas.

completo; el segundo, al estar sujeto a la declaración de impuesto sobre las personas físicas no tiene en cuenta aquellos alquileres en manos de personas jurídicas o personas no sujetas a realizar la declaración por nivel de ingresos; el tercero, no tiene en cuenta los alquileres no declarados lo que también elimina un conjunto relevante de los alquileres reales; por último, aunque el empleo de medianas permite controlar las medidas antes casos atípicos, tiende a mostrar comportamientos erráticos en aquellas zonas en las que tenemos altos niveles de heterogeneidad en la muestra.

2.3.3.3 Índices publicados por el INE

El INE distribuye principalmente tres índices de precios oficiales y uno experimental. Los oficiales son: el índice del precio de la vivienda de compraventa (en adelante IPV); el índice de la vivienda armonizado para Eurostat; el Índice Nacional de la vivienda en alquiler que forma parte del IPC; y el Índice de Precios de la Vivienda en Alquiler (2021b).

El IPV es un índice publicado trimestralmente basado en los precios de compraventa de la vivienda libre[33] registrados por el Consejo General del Notariado. Se estima que la estadística incluye más del 95% de las compraventas totales (INE, 2016a). Se construye como un índice de Laspeyres encadenado con ajuste hedónico en la que las ponderaciones se toman para los dos años naturales anteriores al corriente. El índice aplica dos criterios de estratificación geográfica: nacional o comunidad autónoma y por tipo de vivienda: nueva o segunda mano.

El índice ha usado históricamente dos bases, la primera correspondiente al año 2006 (INE, 2009), y la que está actualmente vigente que se adaptó a la 2015 (INE, 2016a).

El INE además contribuye a Eurostat con el cálculo del HPI[34], que es el índice del precio de la vivienda armonizado, que se realiza de forma simultánea en todos los estados miembros de la UE (Eurostat, 2022) y según la metodología propuesta por la agencia (Eurostat, 2013). Es un índice con base 2015, calculado trimestralmente

Las diferencias metodológicas entre el IPV y el HPI tal y como las describe el INE (2016a) son:

"... El IPV y el HPI se diferencian en dos aspectos técnicos: por un lado, el periodo de referencia de las ponderaciones es el año previo al corriente, en el caso del HPI, mientras que el IPV utiliza los dos años anteriores para su cálculo. Por otro, el HPI incorpora el IVA en el precio de la vivienda nueva y el IPV no lo incluye ..."

[33] Viviendas residenciales que no son ni viviendas sociales o de protección oficial (VPO).
[34] Armonized House Price Index o índice de precios de vivienda armonizado.

El HPI se desglosa sus datos en dos series: el primera, un índice de precios de la vivienda residencial, calculado sobre los datos registrados en las compraventas; y la segunda, denominado OOHPI acrónimo de *Owner Occupied Housing Price Index* (Eurostat, 2017a), cuyo objetivo es representar los costes de comprar, mantener y vivir en una casa en propiedad.

El IPVA se ofrece desglosado zonalmente en cinco niveles: nacional, comunidad autónoma, provincia, municipal (en municipios de más de 10.000 habitantes) y por distritos de capitales de provincia.

Adicionalmente, el INE también incluye datos de precios del alquiler en el IPC[35], a través del Índice Nacional de la Vivienda en Alquiler. Los datos nacionales de la serie están disponibles desde 1961, y desde 2002, cuenta series mensuales desagregadas por provincia[36] . La última modificación de la metodología corresponde a 2017[37] y calcula el precio como una media geométrica de una selección de viviendas para cada provincia. Estos precios proceden de las cuotas mensuales pagadas por los inquilinos que forman el panel de encuestado s.

2.3.3.4 Índices publicados por otras entidades

Además de la administración, los portales inmobiliarios también hacen públicos los datos estadísticos de las viviendas publicadas. En España, las plataformas más utilizadas para la búsqueda de casa son Idealista, Fotocasa, Habitaclia, Pisos.com y Milanuncios.

En diciembre de 2022, según datos de Similarweb[38], se contabilizaron en España un total de 110 millones de páginas vistas en portales inmobiliarios, del que los cuatro portales principales: Idealista, Fotocasa, Habitaclia y Pisos.com, en este orden, acaparaban un 70% del tráfico. En septiembre de 2021, Fotocasa contaba con 65.332 viviendas publicadas en alquiler y 710.375 en venta [39], mientras que Idealista contabilizó 98.987 viviendas en alquiler y 684.073 en venta [40].

Idealista y Fotocasa publican mensualmente indicadores de precio de la evolución del mercado inmobiliario. Los precios de ambas fuentes ofrecen magnitudes ligeramente diferentes, puesto que, Idealista ofrece datos residenciales que incluyen obra nueva y segunda mano, mientras que, Fotocasa calcula las medidas solo sobre pisos en segunda mano. Ambos portales realizan un proceso de

[35] Índice de Precios al consumo
[36] El dato del IPC armonizado se extrae filtrando el subgrupo ECOICOP del alquiler de la vivienda (041).
[37] Aunque la base actual es la del 2021, la última modificación que afectó a los datos de vivienda fue en la metodología con base 2016 (INE, 2017).
[38] Similarweb (2022) es una compañía que ofrece servicios de análisis web, como medidas de volumen de tráfico y de usuarios a sitios de internet.
[39] Datos de anuncios publicados en Idealista (Idealista, 2021) el 12 de septiembre de 2021.
[40] Datos de anuncios publicados en Fotocasa (Fotocasa, 2021) el 12 de septiembre de 2021.

tratamiento de casos atípicos, previamente al cálculo de las series de precios.

Idealista publica el “Informe de Precio Idealista”, que contiene la evolución de los precios publicados de su portal para las viviendas en alquiler y venta. Dispone de un desglose de la información zonal de 6 niveles: barrio, distrito, municipio, provincia, comunidad autónoma y nacional. Las series se calculan como medianas en las zonas utilizadas por el portal, en euros por metro cuadrado / mes, según indica su metodología (Idealista, 2019).

El Índice Inmobiliario Fotocasa tiene una orientación similar, pero incluye solo precios de pisos y áticos de segunda mano anunciados en su portal. Esta fuente ofrece precios en euros por metro cuadrado / mes desde diciembre de 2006, como describe su manual metodológico (Fotocasa, 2017).

2.3.3.5 Índices de precios de la vivienda en Europa y la OCDE

Dado que los índices son estadísticas fundamentales para los responsables de las políticas económicas y monetarias, desde el año 2012, Eurostat exige a los diferentes oficinas de estadística nacionales la creación de índices de la vivienda armonizados.

Para facilitar la homogenización de las metodologías entre agencias estadísticas, Eurostat (2014) desarrolló un manual con recomendaciones para la construcción de índices de precios de la vivienda residencial. Esta guía aconseja el uso de índices hedónicos, con un tratamiento diferenciado entre viviendas unifamiliares y plurifamiliares, considerando las notables diferencias entre las variables disponibles y características constructivas de cada país. Todo ello dificulta la definición de un conjunto canónico de variables y métodos para todos los países de la Unión Europea.

A pesar de las recomendaciones, en la práctica, cada agencia decide la metodología a utilizar, dando lugar a cierta diversidad de criterio como apunta Hill (2018) en su análisis. La Tabla 2.6 resume las técnicas usada por cada país, se puede observar la variedad de técnicas aplicadas, siendo la más popular la basada en medianas estratificadas con ajuste mixto, a pesar de que la agencia europea de estadística recomienda los método hedónicos.

Tabla 2.6. Métodos utilizados para construir los HPI en los países de la UE

Método	**Países**
Revisión de precios	Austria, Bélgica, Finlandia, Hungría, Italia, Letonia, Luxemburgo, Noruega, Eslovenia
Media por características	Rumanía, España
Inputación hedónica	Alemania, Reino Unido
Dummy de tiempo rotativo (RTD)	Croacia, Chipre, Francia, Irlanda, Portugal
Mediana estratificada	Bulgaria, República Checa, Estonia, Islandia, Lituana, Polonia y Eslovaquia
Ratio entre Tasación y Precio de venta (SPAR)	Dinamarca, Países Bajos y Suecia

Fuente: elaboración propia

Aparte de los datos publicados por Eurostat, la OCDE[41] (2018) publica un índice de precios sobre el nominal de precios de la vivienda residencial (IPV Total), para todos sus países miembros, que se desglosa para viviendas nuevas (IPV Nuevo) y de segunda mano (IPV Existente)[42] . Las series de datos se publican trimestrales, excepto para Canadá, Israel, Japón, Corea, Turquía, Brasil, China y Sudáfrica, que publican los datos mensualmente . Casi todos los países publican tres índices de la vivienda (total, segunda mano y nueva), excepto los que se muestran en la Tabla 2.7, que publican uno o dos de ellos .

Tabla 2.7. Índices de precios de vivienda publicados en países de la OCDE

País	**Índice de precios de vivienda (IPV)**
Suiza	Precios de venta de viviendas plurifamiliares (nuevas y existentes) Precios de venta de viviendas unifamiliares (nuevas y existentes)
China	Viviendas plurifamiliares (nuevas) en capitales de provincia
Estados Unidos	Viviendas unifamiliares de segunda mano a nivel nacional
Colombia	Viviendas unifamiliares de segunda mano en áreas urbanas
Australia, Israel	Viviendas nuevas y de segunda mano (todos los tipos) a nivel nacional
Grecia	Viviendas nuevas y de segunda mano (todos los tipos) en áreas urbanas
Corea del Sur	Viviendas de segunda mano (todos los tipos) a nivel nacional

Fuente: elaboración propia

[41] Organización para la Cooperación y el Desarrollo Económicos.

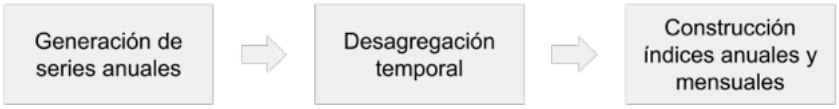
[42] La OCDE distingue tres tipos de índices: “Total” que cubre todas las viviendas (nuevas y segunda mano); “Nuevo” solo para obra nueva; y “Existente” solo para segunda mano. Los tres tipos están disponibles en todos los países, excepto en algunos casos en los que publican uno o dos de ellos.

2.3.4 Desagregación temporal de las series

Dada la metodología utilizada, las series de precios de mercado que se calculan inicialmente son anuales, dado que los registros oficiales de los que proceden también lo son.

Si bien se disponen de modelos de precios de oferta mensuales, con los que se podrían calcular precios de mercado mensuales, el resultado de agregar estas series no coincidirían exactamente con las series de los modelos anuales, por ello se requiere un proceso de conciliación temporal entre los datos anuales y mensuales. Este tarea consiste en desagregar los valores anuales, tomando como referencia las series mensuales creadas con los modelos hedónicos, para su posterior uso en los índices de precios mensuales. La secuencia de etapas en las que consiste el proceso de elaboración del índice, se recogen en la Figura 2.11.

Figura 2.11. Pasos en la generación final de índices



Fuente: elaboración propia.

El proceso de desagregación temporal se define como un método que permite descomponer una magnitud a una escala inferior utilizando un modelo o información auxiliar. En particular, la desagregación temporal de series temporales permite construir series de alta frecuencia (mensual o diaria) a partir de otras de baja frecuencia (anual a trimestral) (Dagum y Cholette, 2006a). Estas técnicas son de uso común en los institutos de estadística nacionales. Por ejemplo, Francia, Italia y otros países europeos, calculan las cifras trimestrales del Producto Interno Bruto (PIB) utilizando métodos de desagregación temporal (Eurostat, 2015; Sax y Steiner, 2013).

Tomando como base la guía de recomendaciones de Eurostat (2015), en su publicación *"Guidelines on temporal disaggregation: benchmarking and reconciliation"*, se indica como plantear un modelo de desagregación temporal usando la información de los niveles de precios originales, previo a la construcción de los índices de precios encadenados.

Existe un gran número de métodos aplicables: los clásicos basados en las primeras diferencias de las series (Cholette y Dagum, 1994; Denton, 1971); los basados en el mantenimiento de la tasa de crecimiento (Causey y Trager, 1981); aquellos basados en regresión (Chow y Lin, 1971; Fernandez, 1981; Litterman, 1983); o

métodos bayesianos (Dagum y Cholette, 2006a; Rojo-García y Sanz-Gómez, 2005; Sayal *et al.*, 2017).

Se ha observado de forma empírica, con los resultados de los índices de alta y baja frecuencia calculados, que las distintas zonas muestran comportamientos muy diferentes en función del submercado de la vivienda que representan. Por lo tanto, se puede asumir la dificultad de encontrar un método único que sea el óptimo para todos los casos.

Por otra parte, las series generadas deben asegurar un nivel de calidad adecuado, solventando los problemas clásicos generados en los procesos de desagregación temporal (Chen y Andrews, 2008; Hood, 2005). En nuestra investigación se aplican tres criterios de calidad:

- Cumplimiento de los requisitos de agregación: se impone que la media de las series desagregadas deben coincidir con el valor de la serie agregada.
- No se deben observar cambios bruscos entre año y año: es frecuente encontrar variaciones bruscas en las series entre los meses de noviembre y febrero.
- La serie estimada no debe mostrar cambios bruscos al inicio o fin del periodo de estudio, habituales en el método Denton[43].

En método propuesto parte del supuesto de que *a priori* se desconoce qué método sería mejor para cada serie, y que el método óptimo podría variar en función de su naturaleza. Un criterio para identificar el método sería a través de estimar la verosimilitud de ofrecer una desagregación de calidad. Esta pauta guarda un paralelismo con el criterio de un experto, que se basaría en seleccionar la serie que se comporte lo más parecido a una serie de referencia, y cuyo proceso ofrezca unas métricas estructurales cercanas al óptimo.

Se ha definido un estimador de máxima verosimilitud propio ($\mathcal{L}$), el cual selecciona la serie cuya medida de verosimilitud θ de los parámetros de calidad sea máxima. Esta medida, para un método de desagregación m y una serie de precios H, se calcularía según la expresión:

$$\mathcal{L}(\theta|\hat{H^m}) = \prod_{c=1}^{n} p(\hat{H_c^m} \mid \theta) \qquad [2.28]$$

donde la verosimilitud se calcula es el producto de probabilidades de un conjunto de parámetros de calidad[44] c para la serie desagregada H con el método m. Con el objeto de facilitar el cálculo, se asume independencia de los sucesos.

[43] El método Denton se describe con detalle en el Anexo 7a del capítulo 7.
[44] Son parámetros calculados según los criterios de calidad mencionados anteriormente

El proceso de selección elige para cada serie temporal, de un total de cinco métodos de desagregación, el método m que hace máxima la verosimilitud $\mathcal{L}$. Los métodos candidatos utilizados son: Chow-Lin, Litterman, Dynamic, Denton-Cholette y Causey-Trager[45].

En un estudio experimental sobre las series generadas por el método hedónico final, desarrollado en el epígrafe 7.4, se demuestra que este proceso de selección de métodos ofrece un conjunto de series de alta frecuencia con una calidad media mayor que si se usara un único método. Por tanto, la hipótesis de utilizar un método a medida de la serie a desagregar resulta acertada.

2.3.5 Construcción de índices finales

El índice mensual se define como un índice encadenado de Fisher, I_{Fo} con base noviembre de 2011. Las ponderaciones utilizadas son el valor de las transacciones y precios en cada periodo t: $p_{i,t}, q_{i,t}$, y las cantidades y precios del periodo base: q_{i0}, p_{i0}. Su cálculo se describe en la siguiente expresión analítica:

$$I_F = \sqrt{\frac{\sum_{i=1}^{n} \hat{p}_{it}\ q_{i0}}{\sum_{i=1}^{n} \hat{p}_{i0}\ q_{i0}} \times \frac{\sum_{i=1}^{n} \hat{p}_{i0}\ q_{it}}{\sum_{i=1}^{n} \hat{p}_{i0}\ q_{i0}}} \qquad [2.29]$$

Se ha utilizado un índice de tipo superlativo (Fisher), con el objetivo de garantizar un mejor cumplimiento con las condiciones ideales exigidas a un índice, que tal y como sugieren Hill (2013) y Eurostat (2014) debe hacerse en la medida de lo posible.

El índice es de tipo encadenado, es decir que la base de cada número índice no es el periodo base sino el periodo inmediatamente anterior. Esta modalidad reduce la influencia de las fluctuaciones en el precio o las cantidades; hace que la magnitud que expresa el índice sea más comparable a lo largo del tiempo; y, adicionalmente, limita la dispersión de los índices de Laspeyres y Paasche (que son la base del de Fisher), (Syed y De Haan, 2017). Sin embargo, puede introducir una ligera deriva (*drift*) en los números índices producidos (Eurostat, 2014; Hill, 2013).

El índice se construye de forma estratificada, de manera que es posible desagregar las magnitudes en función de criterios geográficos o funcionales. El desglose geográfico más profundo es el nivel de barrio, en la ciudad de Madrid, y municipio en el resto de la Comunidad, lo que da lugar a un total de 147 ámbitos geográficos.

[45] Estos métodos se abordan de forma más abundante en el Anexo 7a del capítulo 7.

2.4 Fuentes de información

Esta sección describe los distintos conjuntos de datos utilizados para el desarrollo de la metodología, principalmente fuentes oficiales, como el Censo de Población y Viviendas, la Encuesta de Presupuestos Familiares, Catastro, el Padrón Continuo y diversos indicadores socio-demográficos; a los que se añaden los anuncios de alquiler publicados en el portal de anuncios clasificados idealista. Finalmente, se tratan las problemáticas encontradas sobre los conjuntos de datos y los métodos aplicado para solucionarlas.

Se parte de seis fuentes, recogidas en la Tabla 2.8, que utilizarán en los procesos de construcción de los modelos hedónicos e índices de precios.

Tabla 2.8. Fuentes de información utilizadas

Fuente	Descripción	Motivación
Catastro	Número de parcelas catastrales de cualquier uso	Complementar la información de características de las viviendas de oferta
Censo de Población y Viviendas	Censo de Población y Viviendas desarrollado por el INE en 2011	Representar el colectivo del mercado del alquiler en el periodo base
EPF	Encuesta de Presupuestos Familiares	Permite extrapolar la situación de mercado del año base a años subsiguientes
Idealista	Foto mensual de anuncios en oferta desde noviembre de 2011 a diciembre de 2019	Construir modelos hedónicos de oferta y ser información de apoyo para la construcción del modelo de mercado
Open Street Map	Puntos de Interés y red viaria	Construir características de zona, como los índices de accesibilidad a servicios de las viviendas
Padrón municipal	Población según las divisiones administrativas, series desde 2011 a 2019	Representar la evolución de la población en las unidades geográficas de trabajo

Fuente: elaboración propia

Se usan con dos fuentes de tipo alternativo: los datos del portal inmobiliario Idealista y los datos cartográficos de Open Street Map. Adicionalmente, en el proceso también intervienen otras fuentes menores, como datos de ingresos familiares de la Comunidad de Madrid, atributos socio-demográficos, registros de alquiler vacacional, que no se describen en detalle por su carácter marginal y secundario.

Todas las fuentes se circunscriben geográficamente a la Comunidad de Madrid, para el periodo de tiempo comprendido entre noviembre de 2011 y diciembre de 2019.

2.4.1 Encuesta de presupuestos familiares

La Encuesta de Presupuestos Familiares[46] es una estadística que representa los aspectos clave del gasto de los hogares españoles, cuyos objetivos, según recoge el INE (2006a), son los siguientes:

- Estimación del gasto de consumo anual de los hogares, para el conjunto nacional y para las comunidades autónomas, así como su desglose según diversas variables del hogar.
- Obtención del cambio interanual del total del gasto de consumo, nacional y por comunidad autónoma.
- Estimación del consumo en cantidades físicas para distintos bienes alimenticios.

Además de los tres objetivos principales, destacan por su importancia otros dos vinculados a las necesidades concretas de diversos usuarios de la encuesta: la estimación del gasto como instrumento para la obtención del consumo privado en la Contabilidad Nacional, y la estimación de la estructura de ponderaciones a partir del gasto necesaria para el cálculo del IPC.

Esta estadística se viene desarrollando desde el año 1997, en un primer momento con frecuencia trimestral hasta 2006, cuando se comienza a actualizar anualmente. El tamaño muestral inicial es aproximadamente de 24.000 hogares (INE, 2006a), cada hogar colabora aporta información en dos colaboraciones en años sucesivos.

Los gastos de consumo contenidos en la EPF se refieren tanto al flujo monetario que destina el hogar y cada uno de sus miembros al pago de determinados bienes y servicios (considerados como bienes y servicios de consumo final), como al valor de los consumos efectuados por los hogares en concepto de autoconsumo, autosuministro, salario en especie, comidas gratuitas o bonificadas y alquiler imputado a la vivienda en la que reside el hogar (cuando es propietario de la misma o la tiene cedida gratuita o semi-gratuitamente por otros hogares o instituciones).

El gasto en consumo final de los hogares se registra a precios de adquisición, es decir, al precio que debería pagar efectivamente el comprador por los productos en el momento de la compra y según su precio al contado. Se recoge el importe real de los gastos en bienes y servicios, más todo gasto añadido que hubiera sido provocado por su compra. El gasto en un bien debe registrarse en el momento en que tiene lugar el cambio de propiedad y el gasto en un servicio, en general, cuando se completa la prestación del mismo.

[46] Más información en la web del INE: https://www.ine.es/prensa/epf_prensa.htm

La encuesta se estructura en una serie de capítulos normalizados a nivel europeo, introducidos en el año 2016, y denominada ECOICOP (European Classification of Individual Consumption by Purpose). Esta clasificación, además de ofrecer un mayor desglose de algunas de las parcelas de gasto, permite la comparabilidad con otras estadísticas como el Índice de Precios de Consumo (IPC). También se incorporan cambios en la recogida de la información, como los periodos de anotación en los que se solicitan algunos gastos y los cuestionarios en los que se registran los mismos. La ECOICOP se organiza en doce grupos (INE, 2016b):

1. Alimentos y bebidas no alcohólicas.
2. Bebidas alcohólicas y tabaco.
3. Vestido y calzado.
4. Vivienda, agua, electricidad, gas y otros combustibles.
5. Muebles, artículos del hogar y artículos para el mantenimiento corriente.
6. Sanidad.
7. Transporte.
8. Comunicaciones.
9. Ocio y cultura.
10. Enseñanza.
11. Restaurantes y hoteles.
12. Otros bienes y servicios.

Los doce grupos disponen de un nivel de desglose mayor (códigos de hasta 4 o 5 dígitos[47]), y que recogen la desagregación de los gastos desde un punto de vista funcional. El INE (2016b) desglosa a nivel nacional todas las partidas con mayor profundidad (códigos de hasta 5 dígitos), mientras que los datos relativos a cada comunidad autónoma se desglosan con códigos de 4 dígitos.

La metodología original del 2006 se revisó en 2016, coincidiendo con la normalización europea de categorías y la información procedente del censo de Población 2011. Se procedió al ajuste de la estratificación, subestratificación y afijación [48], así como la renovación parcial del seccionado.

El INE publica los resultados de la encuesta en diferentes formatos: como informe, como estadísticas agregadas y como ficheros de microdatos. En nuestro caso se usan tres ficheros de microdatos descritos a continuación: (1) fichero de hogares, con un registro por cada hogar de la muestra; (2) el fichero de miembros del hogar, que incluye las características de los individuos que componen cada hogar de la

[47] Los códigos numéricos ECOICOP se corresponden con una jerarquía por categorías de gastos. La longitud de los códigos indica el nivel de desglose de la información, siendo el menor desglose los códigos de 2 dígitos, que se corresponden a los grupos, mientras que los códigos de 4 y 5 dígitos contienen información de productos concretos dentro de los grupos.

[48] En un muestreo estratificado, se refiere generalmente a la determinación del número de unidades en la muestra de cada estrato. En el muestreo por conglomerados, se refiere a la decisión sobre el número de conglomerados a seleccionar y el tamaño muestral de cada conglomerado.

muestra; (3) el fichero de gastos, con los valores y categorías de gastos de cada uno de los hogares.

Tabla 2.9. Campos del fichero de la EPF

Campo	Descripción	Valores	Ejemplo
ANNOCON	Fecha construcción edificio	Hace menos de 25 años, Hace 25 ó más años	Hace menos de 25 años
CAPROV	Es capital de provincia Si o No	Sí, No	Sí
CCAA	Código de comunidad autónoma	Numérico o código de elemento	Andalucía
DENSI	Densidad de población del área	Zona densamente poblada, Zona intermedia, Zona diseminada	Zona densamente poblada
GASTOT	Importe total del gasto anual del hogar monetario y no monetario, elevado temporal y poblacionalmente) (para el salario en especie se contabiliza tanto el importe del pago realizado como la bonificación recibida)	Numérico o código de elemento	30075583,26
INTERINPSP	Intervalo de ingresos mensuales netos totales del miembro del hogar	Menos de 500 €, De 500 a menos de 1000 €, De 1000 a menos de 1500 €, De 1500 a menos de 2000 €, De 2000 a menos de 2500 €, De 2500 a menos de 3000 €, 3000 o más €	Menos de 500 €
NHABIT	Número de habitaciones	1 o 2 habitaciones, 3 habitaciones, 4 habitaciones, 5 o más habitaciones	1 o 2 habitaciones
SUPERF	Superficie útil de la vivienda	Desde 35 a 300	106
TAMAMU	Tamaño del municipio	Municipio de 100.000 habitantes o más, Municipio con 50.000 o más y menos 100.000 habitantes, Municipio con 20.000 o más y menos de 50.000 habitantes, Municipio con 10.000 o más y menos de 20.000 habitantes, Municipio con menos de 10.000 habitantes	Municipio de 100.000 habitantes o más

* Las superficies se acotan entre 300 y 35 metros cuadrados, cualquier valor que exceda los extremos se fija al valor máximo y mínimo, por ejemplo, los valores de 350 y 28 se registrarían como 300 y 35 m²n respectivamente.
Fuente: elaboración propia

A partir de los ficheros originales, se han generado dos archivos, uno de gasto y otro de hogares. El fichero de gasto contiene los perfiles de gasto de alquiler de cada hogar (real e imputado), mientras que el de hogares recoge las características del hogar. La estructura de campos de cada uno se detalla en la Tabla 2.9 y la Tabla 2.10.

Tabla 2.10. Campos del fichero de la EPF (continuación)

Campo	Descripción	Valores	Ejemplo
TIPOCASA	Tipo de casa	Chalé o casa grande, Casa media, Casa económica o alojamiento	Chalé o casa grande
TIPOEDIF	Tipo de edificio	Vivienda unifamiliar independiente, Vivienda unifamiliar adosada o pareada, Con menos de 10 viviendas , Con 10 ó más viviendas, Otros (destinado a otros fines o alojamiento fijo)	Vivienda unifamiliar independiente
ZONARES	Tipo de zona residencial	Urbana de lujo, Urbana alta, Urbana media, Urbana inferior, Rural industrial, Rural pesquera, Rural agraria	Urbana de lujo
PESOS	Factor poblacional	Numérico	1257,79
alquiler	Gasto del alquiler anual en euros	Numérico o código de elemento	3737,60
lnGASTOT	Logaritmo del importe total del gasto anual del hogar monetario y no monetario, elevado temporal y poblacionalmente) (para el salario en especie se contabiliza tanto el importe del pago realizado como la bonificación recibida)	Numérico o código de elemento	17,21
factorGASTOT6	Discretización del campo lnGASTOT	Menos.de.15.83, De.15.83.a.16.3, De.16.3.a.16.67, De.16.67.a.17.06, De.17.06.a.17.54, Más.de.17.54	Menos.de.15.83
alquilerm	Gasto del alquiler anual en euros/m²	Numérico o código de elemento	35,26

Fuente: elaboración propia

Se realizan unas mínimas adaptaciones de formato sobre los cambios originales. Se ajusta la variable número de habitaciones (variable *NHABIT*) para adaptarla a los rangos que utiliza el fichero idealista. El importe de alquiler procede de las partidas ECOICOP recogidas por la EPF: *04111: Alquileres reales (vivienda principal)* y *04211: Alquileres imputados a la vivienda en propiedad (vivienda principal)*, en años anteriores a 2016. Para el año 2017 y sucesivos, debido a un cambio de codificación del INE, se intercambian por los códigos anteriores por los nuevos *04110* y *04210*, que se corresponden a las mismas partidas. Sobre el dato de precios, se realiza un proceso de eliminación de los valores más extremos de de la distribución, eliminado todos aquellos valores más allá de 1.5 veces el rango intercuartítilico.

La Tabla 2.11 recoge el nivel de representación de los estratos agrupados por tipo de vivienda y si se encuentran en la capital (se recuerda que la muestra de trabajo se centra en información de la Comunidad de Madrid). Se observa

que el total poblacional se estructura en 1,718 estratos (sin contar el año como dimensión de agrupación), de los cuales 2011 cubre un 23,6% de ellos o 2015 un 24,0%. Se deduce que, de forma general, cada año contiene una cuarta parte aproximadamente de las combinaciones posibles. Se aprecia, además, que los niveles de información, en número de estratos cubiertos al año, son mayores en viviendas plurifamiliares que en unifamiliares. Por otra parte, debe destacarse que el número de estratos totales de las viviendas unifamiliares de la capital es 19, lo que denota la escasa presencia de esta tipología en ese ámbito geográfico.

Tabla 2.11. Nivel de cobertura por estratos en la EPF

Capital	Tipo	Estratos	2011	2012	2013	2014	2015	2016	2017	2018	2019
Todos	Todas	1718	23,6%	23,2%	25,4%	24,1%	24,0%	24,0%	23,2%	20,9%	20,2%
No	Todas	1176	22,0%	21,3%	22,3%	20,3%	22,0%	22,3%	20,9%	18,0%	17,4%
No	Plurifamiliar	875	24,5%	23,2%	24,5%	23,1%	24,1%	23,9%	22,7%	19,4%	20,0%
No	Unifamiliar	301	15,0%	15,6%	15,9%	12,3%	15,9%	17,6%	15,6%	14,0%	10,0%
Sí	Todas	542	27,1%	27,3%	32,3%	32,3%	28,2%	27,7%	28,2%	27,1%	26,2%
Sí	Plurifamiliar	523	27,9%	27,9%	32,3%	33,1%	28,9%	28,1%	28,5%	28,1%	27,0%
Sí	Unifamiliar	19	5,3%	10,5%	31,6%	10,5%	10,5%	15,8%	21,1%	0,0%	5,3%

Fuente: elaboración propia

Si atendemos a lo que representa cada estrato dentro de la población, mediante la variable *PESOS* de la EPF, la Tabla 2.12 muestra que la población (calculada como la suma de los pesos de los agregados de estratos) se concentra en las viviendas de tipo plurifamiliar, con una presencia similar en los ámbitos de la capital y fuera de la capital.

Tabla 2.12. Distribución de pesos poblacionales de la EPF por estratos

Capital	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Sí	Todas	53,3%	52,3%	51,8%	50,7%	49,6%	49,1%	49,3%	48,2%	49,6%
No	Todas	46,7%	47,7%	48,2%	49,3%	50,4%	50,9%	50,7%	51,8%	50,4%
Sí	Plurifamiliar	53,2%	52,0%	50,9%	50,5%	49,4%	48,8%	48,9%	48,2%	49,5%
No	Plurifamiliar	42,0%	42,6%	43,3%	45,5%	44,8%	44,8%	45,3%	46,0%	45,8%
Sí	Unifamiliar	0,1%	0,4%	0,9%	0,3%	0,2%	0,3%	0,5%	0,0%	0,1%
No	Unifamiliar	4,6%	5,0%	4,9%	3,8%	5,6%	6,2%	5,4%	5,7%	4,6%

Fuente: elaboración propia

Por contra, la población de viviendas unifamiliares en la capital es prácticamente inexistente, con valores menores al 1 % (en particular en 2011 su peso era del 0,1 %). En el resto de la provincia tienen mayor presencia, con valores que varían entre el 4 % y el 6 % del total. En general este segmento dispone de una muestra

pequeña y variable, particularmente para el año 2014, con una caída de casi un punto porcentual (de 4,9 a 3,8 %).

2.4.2 Censo de Viviendas y Población

El Censo de Población y Viviendas[49] es una estadística desarrollada por el INE cuyo objetivo es dar a conocer las características de las personas, hogares, edificios y viviendas que existen en España. Comenzó a desarrollarse en el año 1857, y posteriormente, a partir de 1981, se añadió la información de las viviendas. Se publica aproximadamente cada 10 años, y su última publicación se corresponde al año 2011.

En el censo de 2011 se dispone de información tanto de personas que residen en viviendas (ya sean viviendas familiares convencionales o alojamientos) como de las que habitan en establecimientos colectivos (hoteles, residencias, asilos, etcétera).

En nuestro caso, se parte de una extracción de microdatos del censo de 2011 para la Comunidad de Madrid, que contiene un total de 209.449 registros, de los cuales 171 se corresponden a la ciudad de Madrid, y 209.278 a otros 178 a municipios de la Comunidad. Del conjunto de hogares disponibles un total de 22.915 registros (un 11 %) corresponden a hogares en régimen de alquiler.

El desglose de variables disponibles del fichero se describe en la Tabla 2.13. Se aprecia que dispone de un alto grado de detalle en las características físicas de la vivienda, sus suministros, régimen de tenencia y se dispone con el factor de elevación que representa cada observación con respecto a la población total de viviendas. Por último, es importante destacar que en este fichero no dispone del precio de alquiler de la vivienda, por tanto, se utilizará solo para construir los elevadores muestrales de la oferta.

Tabla 2.13. Descripción de campos del censo de viviendas

Descripción	Valores	Valores distintos	Ejemplo
Código de provincia	28	1	28
Código de municipio	Numérico o código de elemento	179	001
Código de barrio	Numérico o código de elemento	129	
Nombre de barrio	Numérico o código de elemento	129	
Tipo de vivienda	Vivienda principal, Vivienda secundaria, Vivienda vacía	3	Vivienda principal

[49] Documentación completa en https://www.ine.es/censos2011_datos/cen11_datos_resultados.htm

Régimen de tenencia	Propia, por compra, totalmente pagada, Propia, por compra, con pagos pendientes (hipotecas), Propia por herencia o donación, Alquilada, Cedida gratis o a bajo precio (por otro hogar, pagada por la empresa...), Otra forma	6	Propia, por compra, totalmente pagada
Calefacción	Colectiva o central, Individual, No tiene calefacción pero sí algún aparato que permite calentar, No tiene calefacción	4	Colectiva o central
Cuarto de aseo con inodoro	Si, No	2	Si
Baño o ducha	Si, No	2	Si
Acceso a Internet	Si, No	2	Si
Sistema de suministro de agua	Agua corriente por abastecimiento público, Agua corriente por abastecimiento privado o particular del edificio, No tiene agua corriente	3	Agua corriente por abastecimiento público
Superficie útil	Numérico o código de elemento	458	
Número de habitaciones	Numérico o código de elemento	25	
Número de plantas sobre rasante	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más	10	1
Número de plantas bajo rasante	, 0, 1, 2, 3	5	
Tipo de edificio	Destinado principal o exclusivamente a viviendas, Destinado a otros fines	2	Destinado principal o exclusivamente a viviendas
Número de inmuebles	Numérico o código de elemento	18	1
Año de construcción	Numérico o código de elemento	19	Antes 1900
Estado del edificio	Ruinoso, Malo, Deficiente, Bueno	4	Ruinoso
Ascensor en el edificio	Si, No	2	Si
Accesibilidad del edificio	Si, No	2	Si
Garaje	Si, No	2	Si
Número de plazas de garaje	1, 2, 3 a 5, 6 a 10, 11 a 20, 21 a 50, Más de 51	7	1
Gas	Si, No	2	Si

Tendido telefónico	Si, No	2	Si
Agua caliente central	Si, No	2	Si
Evacuación de aguas residuales	Alcantarillado, Otro tipo, No tiene sistema de evacuación de aguas residuales	3	Alcantarillado
Factor de elevación		0	

2.4.3 Anuncios en el portal Idealista

Para conocer en detalle la oferta, se utiliza una base de datos proporcionada por el portal inmobiliario Idealista para la Comunidad de Madrid[50], procedente de una extracción de datos mensual de sus anuncios de alquiler, publicados en el periodo entre 2011 y 2019.

Se cuenta con dos tipos de viviendas residenciales, las unifamiliares cuyos subtipos en el fichero son: pareados, adosados y viviendas aisladas, y las plurifamiliares compuestas por los subtipos: pisos, estudios, dúplex y áticos.

Cada registro de la base de datos corresponde a un anuncio publicado en el portal Idealista para un mes dado. Una observación contiene una lista de atributos físicos de la vivienda como: sus metros cuadrados; el número de habitaciones; el equipamiento; su localización como coordenadas geográficas; e información relativa al nivel de visitas y contactos que recibe cada anuncio.

Los campos difieren ligeramente si se trata de vivienda unifamiliar o plurifamiliar, el listado de campos para ambas se muestra en la Tabla 2.14.

Tabla 2.14. Diccionario de variables Idealista

Variable	Descripción	Tipo
AMENITYID	Tipo de instalaciones de la finca	Edificio
BUILTTYPEID	Nuevo o segunda mano	Edificio
CHALETTYPEID	Tipo de inmueble unifamiliar	Edificio
FLOOR_POSITION	Posición del piso dentro del edificio plantas superiores, medias o inferiores	Edificio
HASDOORMAN	Tiene portero	Edificio
HASGARDEN	Tiene jardín	Edificio
HASLIFT	Tiene ascensor	Edificio
HASSWIMMINGPOOL	Tiene piscina	Edificio

[50] Accesible a través de su web en http://www.idealista.com

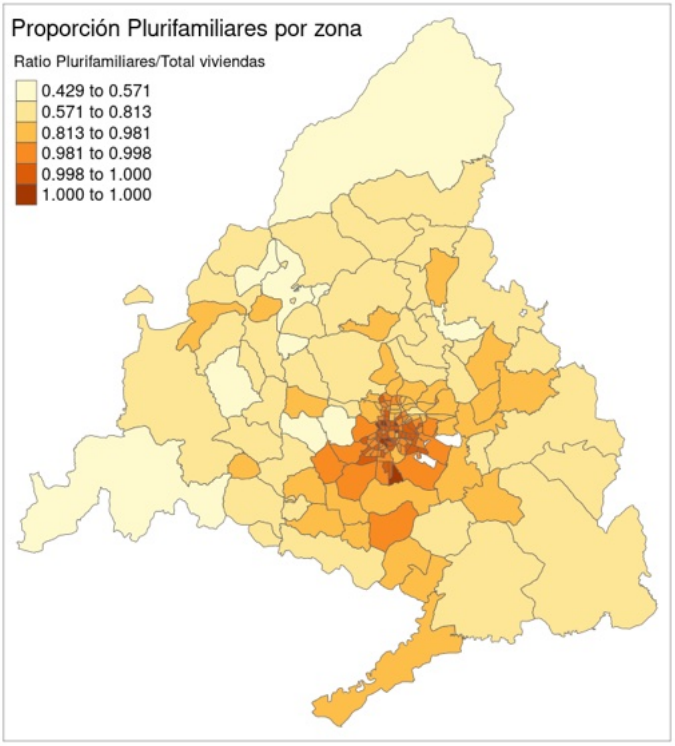
MAXBUILDINGFLOOR	Número de pisos en edificio	Edificio
BATHNUMBER	Número de baños	Estructura
BEDROOMNUMBER	Número de dormitorios	Estructura
CONSTRUCTEDAREA	Superficie total en metros cuadrados	Estructura
ENERGYCERTIFICATIONID	Código de certificado energético	Estructura
FLATLOCATION	Indica si el piso es interior o exterior	Estructura
FLOOR	Planta en la que está el inmueble	Estructura
GARAGETYPEID	Tipo de garaje	Estructura
HASAIRCONDITIONING	Tiene aire acondicionado	Estructura
HASANNEX	El piso tiene anejos (garaje o trastero)	Estructura
HASBALCONY	Tiene balcón	Estructura
HASBOXROOM	Tiene almacenamiento / Trastero	Estructura
HASEASTORIENTATION	Está orientado al este	Estructura
HASNORTHORIENTATION	Está orientado al norte	Estructura
HASPARKINGSPACE	Tamaño del garaje	Estructura
HASSOUTHORIENTATION	Está orientado al sur	Estructura
HASTERRACE	Tiene terrazas	Estructura
HASWARDROBE	Tiene armarios empotrados	Estructura
HASWESTORIENTATION	Está orientado al oeste	Estructura
ISDUPLEX	Es un duplex	Estructura
ISPENTHOUSE	Es un ático	Estructura
ISSTUDIO	Es un estudio	Estructura
PLOTOFLAND	Tamaño de la parcela unifamiliar	Estructura
ROOMNUMBER	Número de habitaciones	Estructura
USABLEAREA	Área útil	Estructura
PERIOD	Código de mes ordinal de 1 a 12, asignado al mes en 2018 cuando se tomó la observación	Fecha
CHANNELID	Canal de venta del inmueble: 1 agente inmobiliario, 2 inmuebles de propiedad bancaria y 3 vendedor individual.	Mercado
LEADS	Número de contactos con el propietario (mensajes) a través de la página web	Mercado
LEADS_RESIDENTIAL	Número de contactos medios en la zona idealista	Mercado
ONMARKET_RENT	Número de inmuebles en alquiler en la zona	Mercado
ONMARKET_SALE	Número de inmuebles en venta en la zona	Mercado
PRICE	Precio/mes en euros	Mercado

RENTSALE_RATIO	Proporción de número de inmuebles en alquiler versus en compra (en oferta)	Mercado
TOTALLISTINGVIEWS	Apariciones del anuncio en listados de búsqueda	Mercado
TOTALVIEWS	Vistas en de la ficha de detalle del anuncio	Mercado
UNITPRICE	Precio/mes en euros por metro cuadrado útil	Mercado

Como en el caso de la EPF la vivienda unifamiliar tiene una menor presencia, en el caso de Idealista la media mensual de anuncios unifamiliares de de 3.661, mientras que la de plurifamiliares es 36.869, que representa una proporción de 10,07 viviendas plurifamiliares por cada unifamiliar (lo esperado en una zona tan urbana y densamente poblada como es la Comunidad de Madrid).

Geográficamente, la capital y su entorno más cercano contiene principalmente viviendas plurifamiliares, mientras que, las zonas periféricas cuentan con una mayor proporción de unifamiliares (Figura 2.12).

Figura 2.12. Proporción de viviendas de tipo plurifamiliar sobre el total

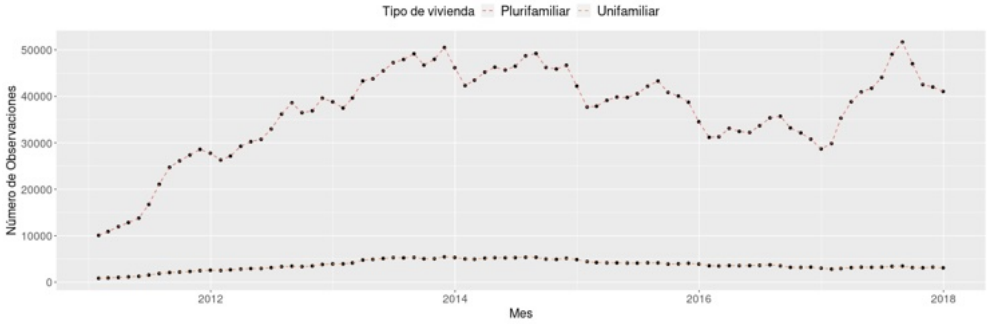


Fuente: elaboración propia.

La Figura 2.13 muestra cierta fluctuación del número de anuncios activos a lo largo del tiempo, con un crecimiento sostenido entre 2011 a 2014, y un decrecimiento hasta final de 2016, con una vuelta a la senda de crecimiento en 2017. Estos efectos se deben a tanto al comportamiento del mercado inmobiliario

como a oscilaciones en la cuota de contenidos del portal.

Figura 2.13. Número de anuncios por tipos, frecuencia mensual



Fuente: elaboración propia.

2.4.4 Open Street Map

Open Street Map (a continuación OSM) es un proyecto colaborativo para crear un mapa del mundo editable y gratuito (OpenStreetMap, 2017), disponible bajo una licencia de base de datos abiertos. La creación de contenidos y el crecimiento de OSM se han visto motivados por las restricciones sobre el uso o la disponibilidad de datos cartográficos en gran parte del mundo, y la adopción generalizada de los dispositivos móviles y navegadores GPS.

Figura 2.14. Ejemplo de topología y red viaria basada en Open Street Map

Fuente: elaboración propia basada en cartografía de Open Street Map (2022).

El proyecto fue creado en el Reino Unido en año 2004 y está respaldado por la *Open Street Map Foundation*[51]. La iniciativa se inspiró en el modelo de la Wikipedia, y se inició con la apertura de una serie de mapas de partes del mundo. Desde entonces, ha aumentado a más de dos millones de usuarios, que contribuyen

[51] Organización sin ánimo de lucro registrada cuyo fin es el desarrollo y disponibilización de información geoespacial gratuita y reutilizable.

con contenidos a través de cargas manuales, dispositivos GPS, fotografías aéreas y otros medios.

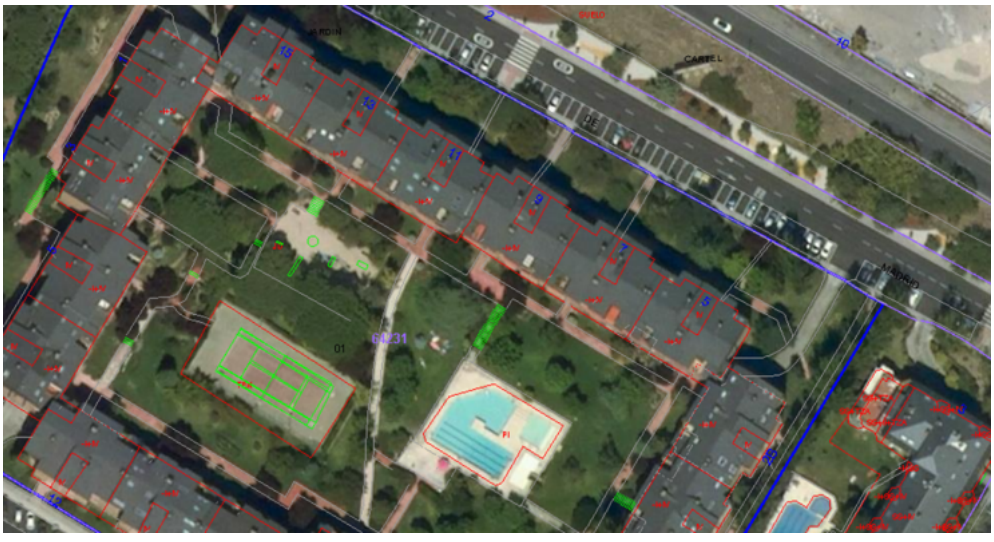
En nuestro caso, se toma la información de OSM procesada por la empresa GeoFabrik[52], aplicada a:

- Extracción de puntos de interés (Point of Interest o POI en inglés), que son ubicaciones en las que se encuentra algún establecimiento útil o representativo (por ejemplo, un restaurante, un hotel, etcétera). Cada uno de ellos asociado a una clasificación de tipos, un nombre del elemento y sus coordenadas geográficas
- Información viaria, necesaria para la construcción de la topología de la red de transporte[53], sobre la que se calculan los tiempos de desplazamiento entre localizaciones, isócronas e isodistancias, como muestra la Figura 2.14.

2.4.5 Catastro

El catastro inmobiliario es un registro administrativo del Estado en que recoge los bienes inmuebles rústicos, urbanos y de características especiales. Es un registro estadístico cuya finalidad es determinar la extensión y valor de cualquier demarcación geográfica, cuyo objeto es ser un instrumento sobre el que se determinan los impuestos sobre bienes inmuebles (Romero *et al.*, 2014). En el caso español, los catastros son responsabilidad municipal, y es la Dirección General del Catastro quien consolida esta información a ámbito nacional (a excepción de las provincias vascas y Navarra).

Figura 2.15. Ejemplo de cartografía catastral sobre ortofoto del PNOA



Fuente: elaboración propia basada en cartografía del PNOA (2022).

[52] https://www.geofabrik.de
[53] En nuestro caso se ha usado la extensión *pgRouting*, desarrollada por Obe y Hsu (2011).

El dato utilizado corresponde a una extracción de la Sede Electrónica del Catastro[54] para la comunidad de Madrid con fecha Diciembre de 2019, (Direccion General del Catastro, 2022).

Los elementos catastrales desde un punto de vista funcional son:

- Parcela catastral o finca catastral: contiene una serie de inmuebles y elementos comunes. Su destino principal [55] puede ser residencial, industrial, público, entre otros (Figura 2.15).
- Inmueble: cada uno de los elementos que componen la finca y con un aprovechamiento determinado: residencial, comercial, deportivo, religioso, etcétera.
- Elementos comunes: elementos que componen la finca junto a los inmuebles, pero de uso compartido entre los propietarios de la finca.

El inmueble y los elementos comunes cuentan con un desglose constructivo denominado “construcción”, el cual representa la superficie construida asociada a un inmueble o un elemento común en su parcela catastral. En la base de datos se registra, además, una relación entre cada inmueble y todos los elementos comunes que contiene. Cada vivienda cuenta con una referencia catastral única de 20 dígitos que se compone de una parte común por finca, de longitud 14 dígitos, y una específica de inmueble o elementos comunes, de seis posiciones alfanuméricas.

Los atributos catastrales utilizados a nivel de finca son: 1) año de construcción o de última gran reforma; 2) calidad constructiva de la finca clasificada con 12 grados, desde muy buena construcción a mala construcción; 3) número de número de inmuebles de la finca; 4) la altura total de la finca y 5) coordenadas geográficas del centroide[56] de la finca. Se omiten atributos como la disponibilidad de ciertas instalaciones en la finca como jardín, piscina, u otras, porque ya están informados en los registros procedentes del portal.

A nivel de inmueble, se dispone de la dirección completa, la superficie total, la cuota de participación sobre elementos comunes de la finca, y la lista de construcciones, con sus respectivos metros cuadrados y uso destino[57].

Como se observa en la Tabla 2.15, el uso habitual en los tipos residenciales es la vivienda plurifamiliar, con una gran diferencia en número con el tipo unifamiliar.

[54] https://www.catastro.meh.es
[55] Se corresponde a los metros construidos para uso destino mayoritario de la finca catastral.
[56] Centro geométrico de un polígono.
[57] El uso destino catastral es un código de tres letras que indica que uso tiene el suelo, existen diversos usos entre los que se encuentra, residencial, comercial, deportivo, etcétera.

Tabla 2.15. Número de elementos catastrales en la Comunidad de Madrid

Elemento	Madrid	Resto CAM	Todas zonas
Finca	319.718	1.192.972	1.512.690
Construccion	3.050.967	4.270.251	7.321.218
Construccion Residencial	1.712.761	1.988.898	3.701.659

Fuente: elaboración propia

A nivel geográfico, existe una mayor presencia relativa de las vivienda plurifamiliares en la capital, dónde el 97% de los inmuebles son pisos (Tabla 2.16).

Tabla 2.16. Distribución de viviendas de tipo residencial

Ámbito	Plurifamiliar	Unifamiliar
Todas las Zonas	85%	15%
Madrid	97%	3%
Resto CAM	72%	28%

Fuente: elaboración propia

2.4.6 Detección y corrección de errores

Gran parte de las fuentes con las que se trabaja ya han sido tratadas en origen, como la EPF, el censo o el catastro. Sin embargo, el dato procedente de un portal inmobiliario tiene varias peculiaridades que obligan a realizar un tratamiento específico (Loberto *et al.*, 2018): en primer lugar, se ha generado por usuarios del servicio y cuenta con un grado de subjetividad, disparidad de criterios a la hora de interpretarlos, y posibles errores de entrada; por otro lado, porque se producen prácticas fraudulentas por parte de ciertos profesionales que distorsionan la muestra; en tercer lugar, el portal ha ido transformándose a lo largo del tiempo (inicialmente solo como página web en PC, después como página para teléfonos móviles y en los últimos años se ha hecho accesible a través también de una aplicación para dispositivos móviles), lo que ha dado lugar a modificaciones en la forma en la que se registra la información y el significado de los campos.

Históricamente, las guías oficiales de las oficinas de estadísticas se han centrado en el tratamiento de fuentes estadísticas primarias, por ejemplo, el Informe de Calidad para la Encuesta Social Europea (2014). Sin embargo, el uso de las técnicas habituales para control de calidad estadístico no es directamente extrapolable a grandes fuentes de datos (Anenberg y Laufer, 2017). Debido al creciente interés en incorporar fuentes alternativas (Biancotti *et al.*, 2020),

en los últimos años estas guías se han adaptado para incluir recomendaciones acerca del tratamiento de fuentes secundarias de información no estadística (por ejemplo, datos de Internet), como UNECE (2015), Struijs y Daas (2014) o la guía de Eurostat (2017b) para la incorporación de fuentes de tipo *"Big Data"*.

Aún cuando los datos que se muestran en la página se supervisan por los equipos de control de calidad del portal, no se controlan todos los errores y el significado de los campos están sujetos a interpretaciones por parte de los usuarios[58]. Para evitar dichos inconvenientes la metodología desarrolla un tratamiento previo de la información, consistente en tres pasos:

- De-duplicación: en el caso de que un mismo inmueble cuente con varios anuncios se toma solo una sola instancia.
- Eliminación de datos atípicos (outliers): consiste en la eliminación de todos aquellos valores considerados muy infrecuentes.
- Imputación de atributos constructivos: para completar los campos esenciales del inmueble o la finca necesarios en los procesos de calibración, como son la calidad constructiva, el número de inmuebles y altura en la finca catastral, y el año de construcción.

En estadística, un valor atípico, es un dato que difiere significativamente del resto de observaciones, y que no necesariamente se trata de un error de medida. Estos valores pueden afectar de manera significativa a ciertos parámetros de la distribución, sobre todo a aquellos no robustos como la media o la varianza. Cuando hablamos de valores atípicos atendemos a cualquier dato que no se encuentre dentro de los valores normotípicos, cuya presencia puede deberse a la existencia de errores en la base de datos o por valores válidos pero muy extremos, que no representan a generalidad de los individuos de la población.

Otro fenómeno que se trata en esta sección será la de imputación de valores ausentes y erróneos, debido a la importancia de algunos campos con un alto grado de información ausente, como los campos de área útil o año de construcción, y que son esenciales en los modelos de precios y en la ponderación de poblaciones. El motivo de tratarlos se debe al potencial impacto que tienen este tipo de valores en la calidad de los modelos a desarrollar (Rubin, 1976).

2.4.6.1 Revisión de técnicas para el tratamiento de valores atípicos

En la literatura existen numerosas definiciones de outlier, como *"[..] Observaciones o medidas sospechosas porque son mucho más pequeñas o grandes que la gran mayoría de las observaciones [...]"* (Cousineau y Chartier, 2010) o *"[..]*

[58] Por ejemplo, el campo "número de habitaciones" podría interpretarse como número de huecos o número de dormitorios.

observación que cae fuera de los parámetros normales de una observación [. . .]" (Jarrell, 1992; Stevens, 1984). Por otro lado Hawkings (1980) en su libro *"Identificación de outliers"*, lo describe como una observación que *"[..] se desvía tanto del resto de observaciones que la hace sospechosa de haberse generado con un mecanismo diferente [..]"*. En otros casos los definen simplemente como valores que son *"[..] dudosos a ojos del investigador [..]"* (Dixon, 1950) o *"[..] contaminantes [..]"* (Wainer, 1976). En el caso de Wainer, se introduce el concepto de *"fringelier"* [59] refiriéndose a *"[..] sucesos inusuales que ocurren más a menudo de lo que deberían [..]"*, por ejemplo, las observaciones que se ubican cerca de tres veces la desviación estándar y que pueden tener una fuerte influencia en la estimación de parámetros, pero que no estar lo suficientemente lejos del centro de la distribución no se consideran atípicos.

Los valores atípicos pueden deberse a la variabilidad de la medición o errores experimentales o de codificación, como en el caso de la información de los portales. Al proceder de información introducida por los usuarios están sujetos a estos errores.

Los valores infrecuentes pueden distorsionar los parámetros en los estadísticos estimados, tanto cuando se usan pruebas paramétricas y/o no paramétricas, como explica Zimmerman (2010; 1995, 1998). Sin embargo, no siempre se realiza una limpieza adecuada en los proyectos de investigación, en particular, podemos ver en el caso de la revisión realizada por Osborne (2001) en el campo de la psicología educativa, donde son pocos los investigadores que controlan estas anomalías.

Una pequeña proporción de casos pueden afectar enormemente los resultados, siendo siempre conveniente su eliminación después de un análisis previo como recomiendan Osborne y Overbay (2004). Como recogen los autores, los efectos principales en los análisis estadísticos son: en primer lugar, el incremento en la varianza y la reducción del poder de los test estadísticos; en segundo lugar, si no están distribuidos de forma aleatoria pueden reducir la normalidad de la distribución (y por tanto el incumplimiento de ciertas asunciones requeridas en algunos procesos), lo que se puede resumir en una alteración de las posibilidades de cometer errores de tipo I [60] y tipo II [61]; y en tercer lugar, la presencia de valores atípicos pueden sesgar seriamente ciertos análisis (Schwager y Margolin, 1982).

El origen de los valores atípicos puede ser múltiple, Anscombe (1960) los clasifica en dos orígenes: los que proceden de errores en los datos o lo que

[59] La palabra *"fringelier"* se puede traducir como elemento que cae en el extremo de un área, procede de la palabra inglesa *"fringe"* que significa el borde exterior de un zona.

[60] El error de tipo I o falso positivo, es el error que se comete cuando el investigador rechaza la hipótesis nula

[61] El error de tipo II o falso negativo, se comete cuando el investigador no rechaza la hipótesis nula siendo esta falsa en la población

proceden de la variabilidad inherente de los datos. No todos los outliers serán posibles contaminantes, ni todos los que se puedan marcar como atípicos serán verdaderamente atípicos (Barnett y Lewis, 1984). Si nos referimos a la clasificación de atípicos que propone Osborne y Overbay (2004) tendríamos cinco categorías:

- Outliers procedentes de errores humanos en la entrada de los datos o los procesos de registros de la información.
- Outliers por errores creados de forma intencionada, como pueden ser respuestas a encuestas en las que los entrevistados tienen un claro interés en sesgar los resultados.
- Outliers por errores en la propia selección de la muestra.
- Outliers por fallos de estandarización o calibración.
- Outliers por fallos en la asunción de la distribución de la muestra, debidos a que se asume una distribución para la población que no se corresponde con la real (por ejemplo asumir que la distribución de la población es normal cuando no lo es).

Existe cierta controversia sobre eliminar o no estos valores. Asimismo, Barnett (1984) considera que *"[...] es de sentido común su eliminación [...]"*. Autores como Judd, McClelland y Ryan (2011) proponen la eliminación del valor atípico aún cuando este sea un valor legítimo, ya que permiten la estimación más correcta de los parámetros de la población. Mientras que otros autores son más reacios a su eliminación (Orr *et al.*, 1991), o bien que solo se eliminen cuando el dato sea erróneo (Cousineau y Chartier, 2010).

En nuestra opinión, en procesos como este, con un gran volumen de observaciones, es prácticamente inviable una revisión pormenorizada de aquellos casos que aún siendo clasificados como atípicos puedan ser una observación legítima, por lo que tenderemos a la eliminación de estas observaciones. En todo caso, estudiará en profundidad cada una de las subpoblaciones clasificadas como atípicas, por si estas fueran de algún interés para futuros análisis. En este sentido, existen alternativas a la eliminación de los registros como el uso de técnicas estadísticas robustas, en las que los outliers *"[...] no representan un inconveniente o bien son robustos ante la presencia de outliers [...]"* (Barnett y Lewis, 1984).

Cousineu (2010) clasifica los métodos de tratamiento de outliers en dos tipos:

- Aproximación univariante: en la que se realiza la clasificación con una sola variable, como por ejemplo, eliminar los valores fuera de 2 veces el rango intercuartílico.
- Aproximación multivariante: en la que se realiza la clasificación con varias

variables. Un caso particular de ella es la bivariante, en la que intervienen dos variables.

Entre las técnicas de detección univariante podemos destacar el test de Grubbs (Tietjen y Moore, 1972), el Dixon (Miller, 1993) o el de Tukey (1953). El principal inconveniente de estos procesos es que exigen que la distribución de la variable sea normal. A este respecto, es bastante habitual en estudios sobre los precios de la vivienda asumir una forma normal logarítmica, aunque lo más apropiado según algunos autores, como Ohnishi (2011), es usar una transformación potencial, por ejemplo Box-Cox, para trabajar con una distribución lo más normal posible.

Por otro lado, una vez se detectan los valores atípicos se pueden realizar varias acciones con ellos:

- La eliminación de la observación. Esta opción no siempre es posible en muestras pequeñas y debe realizarse de forma que no se introduzcan sesgos importantes en la nueva población.
- La imputación del valor. Esta acción se aplica también a observaciones ausentes o *"missing"*. La imputación puede ser a través de enfoques paramétricos como el uso de medias, medianas o moda, o basados instancias similares, como puede ser la técnica *"hot deck imputation"*[62] (Andridge y Little, 2010).
- La limitación del valor (en inglés *"capping"*): se asigna el valor máximo o mínimo aceptable para la variable, un caso de esta técnica es la *winsorización* [63].
- Predecir el valor. Se usa un modelo predictivo para calcular el valor más probable de la variable, usando el resto de campos del registro.

Ciertos procesos como el de la imputación de valores basados en registros similares o la predicción, pueden ser costosos ante conjuntos de datos grandes. Por ese motivo, en ocasiones, se aplica un conjunto combinado de enfoques univariantes y multivariantes sencillos. Estas técnicas estadísticas ofrecen un buen balance entre precisión y coste de cálculo. Además de las aproximaciones mencionadas, también existen métodos basados en modelos de aprendizaje no supervisado, cuyo principio se fundamenta en identificar los casos menos plausibles en la distribución natural de las observaciones de la muestra (Wang *et al.*, 2021).

[62] Esta técnica imputa los atributos de una instancia en función de los atributos de registros similares, como indica el autor se basa en un modelo donante-receptor, en el que se asegura que el receptor recibe la donación de atributos de el individuo más compatible por similitud

[63] En la técnica de winsorización se establecen un valor máximo o mínimo para la población, o asociados a ciertos cuantiles de la distribución, y se asignan estos valores si el valor original es más extremo

2.4.6.2 Tratamiento de duplicados y valores atípicos

Un mismo propietario puede trabajar con varias inmobiliarias a la vez cuando intenta vender su piso, por lo tanto, si todas publican un anuncio para esta vivienda, será posible encontrar registros duplicados en la base de datos a nivel de inmueble.

Para mitigar esta situación, se ha utilizado el algoritmo desarrollado por Idealista que identifica qué anuncios están duplicados dentro del portal. Este proceso se basa en identificar todos los anuncios, en la misma ubicación geográfica, que tienen un alto grado de parecido utilizando mediante una medida de distancia denominada similitud combinada sim, definida como:

$$sim_{A,B} = f(F_A, F_B) \quad [2.30]$$

donde sim es el resultado de aplicar una función f que estima un valor entre 0 (diferente) a 1 (exactamente igual) sobre una serie de características F, para los anuncios A y B. Estas características son: imágenes, descripción del anuncio en texto, atributos básicos de superficie y precio. Para cada par de anuncios, cuando la similitud supera un umbral determinado, se considera que ambos se corresponden a la misma vivienda. El proceso agrupa todas los anuncios de cada vivienda y se selecciona el primer registro publicado como representante, el resto se eliminan de la muestra.

Para la determinación de valores atípicos, se ha decidido aplicar un conjunto de criterios estadísticos básicos junto con alguna regla experta. Se han descartado las modalidades multivariante complejas, por su alto coste en tiempo y proceso en este conjunto de datos tan extenso. Por ello, se utiliza una técnica de eliminación de valores anómalos mediante la combinación de cuatro criterios univariantes y bivariantes: 1) precio por metro cuadrado; 2) ratio superficie sobre numero de habitaciones; 3) año de construcción; 4) tamaño de la finca en el caso de inmuebles unifamiliares.

Se usa el método denominado de Tukey o Boxplot (Ben-Gal, 2005) por su eficiencia y porque el número de variables a controlar es relativamente bajo, descartando por tanto métodos basados en similitudes, o el propio test de Grubbs, al ser un método iterativo que sobre un gran volumen de datos requiere una gran cantidad de tiempo de proceso, no mejorando los resultados.

Este método define que los limites que puede tomar una variable son una razón de su rango intercuartílico. Se definen dos barreras sobre el primer y tercer cuartil, la cercana (“*near*”) como 1,5 veces el rango intercuartilico], y la lejana (“*far out*”),

3 veces el rango intercuartilico. Para el caso de la barrera cercana los límites se calcularían con la siguiente expresión analítica:

$$[\ Q_1 - 1,5 \cdot (Q_3 - Q_1),\ Q_3 + 1,5 \cdot (Q_3 - Q_1)\ ] \qquad [2.31]$$

Donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente.

El método de Tukey es sencillo de aplicar y tiene un buen rendimiento sobre distribuciones sesgadas y asimétricas, aunque existen adaptaciones, como las propuestas por Hubert y Vandervieren (2008), que mejoran las debilidades del método original.

En todo caso, es importante aclarar que se ha evaluado el uso de la técnica *"Local Outlier Factor"* (He *et al.*, 2003) para estimar su posible aplicación en esta investigación. Este método se basa en estimar la distancia de cada instancia con respecto a sus vecinos en características. Se calcula una puntuación para cada instancia, que representa un índice de densidad con respecto a los K vecinos mas cercanos. Los resultados obtenidos muestran que su coste computacional no compensa, pues se observa que apenas logra detectar casos no detectados por la aproximación combinada de medidas univariantes y bivariantes.

Tabla 2.17. Muestra y tasa de outliers

Año	Anuncios	Unifamiliares	Plurifamiliares	Eliminados
2011	203.671	172.895	30.776	17%
2012	340.357	292.681	47.676	20%
2013	420.682	362.459	58.223	22%
2014	484.672	417.386	67.286	21%
2015	513.484	436.353	77.131	22%
2016	546.295	457.796	88.499	24%
2017	555.052	465.700	89.352	24%
2018	634.767	544.896	89.871	25%
2019	711.450	619.990	91.460	25%

Fuente: elaboración propia

Se observa que la proporción de registros descartados por criterios de anomalías o valores extremos varían a lo largo de los periodos. Podemos destacar una tasa creciente de anuncios eliminados desde el 17% en 2011 al 15% en 2019, como muestra la Tabla 2.17. Existen dos motivos para ello: el primero, se debe a que la tasa de anuncios activos, pero sin visibilidad suficiente en Idealista, aumenta ligeramente a lo largo del tiempo [64]; el segundo, a que el fichero contiene fotos

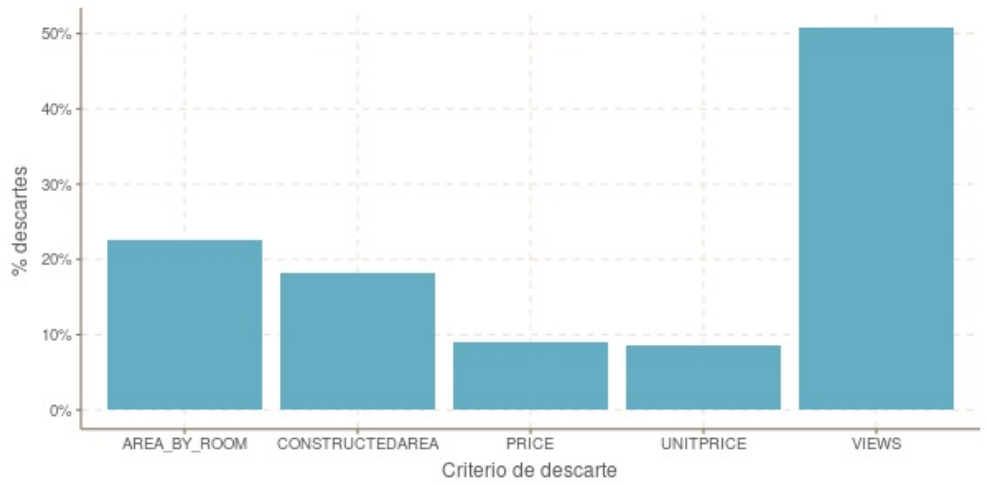
[64] Es importante destacar que la extracción se realiza sobre la base de Idealista y no es un captura

mensuales de anuncios publicados, por tanto si una misma vivienda atípica está publicada por ejemplo en enero y febrero, el mismo anuncio se considerarían como dos registros atípicos.

Como se describía anteriormente, la identificación de anómalos se evalúa con criterio de barreras de Tukey sobre las variables normalizadas. Al ser todas las variables positivas, se han aplicado las transformaciones potenciales de Box-Cox (Sakia, 1992). Para controlar el nivel de normalidad de las variables transformadas se ha utilizado el test de normalidad propuesto por Jarque y Bera (1980).

En la Figura 2.16 se muestran los criterios más aplicados en los registros descartados. El criterio de descarte más común es la eliminación por *VIEWS*[65] (por no tener que generar suficientes visualizaciones del anuncio en el portal). Estos anuncios se corresponden a anuncios con baja demanda publicados en canales secundarios, denominados como *"microsites"*[66] y difundidos en las páginas web de algunas agencias inmobiliarias que son además clientes de idealista .

Figura 2.16. Proporción de registros descartados según criterio



Fuente: elaboración propia.

Adicionalmente, se aplican otros dos criterios, de carácter experto, para filtrar los anuncios:

- La superficie construida debe ser mayor a 35 m^2, en base a la normativa española[67] que exige un número de metros mínimos útiles de vivienda.

del contenido visible en internet. Según nos confirma la empresa, la base comprende tanto anuncios visibles en idealista.com y anuncios menos visibles o sitios de menor afluencia

[65] Este atributo indica el número de veces que un usuario ha visto este anuncio.

[66] Un microsite en terminología de idealista, es un portal explotado por un tercero que se muestra contenido de idealista, en general este tipo de sitios tienen un nivel de visibilidad muy inferior al del portal principal.

[67] Orden de 29 de febrero de (1944) [Ministerio de la Gobernación]. Por la que se determinan las condiciones higiénicas mínimas que han de reunir las viviendas.

- El anuncio debe haber recibido al menos 10 visualizaciones al mes, por tanto se eliminan aquellos inmuebles que no tengan un mínimo nivel de demanda.
- El año de construcción del inmueble, según catastro, debe ser anterior o igual al año informado por el anunciante.

La superficie mínima útil de una vivienda tiene que ser en todos los casos mayor a 36 metros cuadrados.

A modo de resumen, la Tabla 2.18 muestra los rangos de máximos y mínimos de las variables clave utilizados como barreras Tukey para eliminación de atípicos. Se observa que estos valores no se corresponden con valores normotípicos de los inmuebles, por ejemplo, que la proporción del área útil por estancia (AREA_BY_ROOM) sea 89,71 m² en pisos, y 255,22 m² para unifamiliares.

Tabla 2.18. Valores máximos y minimos aceptables por variable y tipo de vivienda

Tipo	Variable	Mínimo	Máximo
Plurifamiliar	AREA_BY_ROOM	7,12	89,71
Plurifamiliar	UNITPRICE	2,89	47,30
Unifamiliar	AREA_BY_ROOM	6,03	255,22
Unifamiliar	UNITPRICE	1,00	31,05

Fuente: elaboración propia

La presencia de datos ausentes plantea dificultades en el desarrollo de los modelos, como pone de manifiesto Rubin (1976), que desarrolló un modelo de inferencia de datos ausentes sobre información incompleta, que aún a día de hoy está en uso. Sin embargo, se pueden encontrar distintos planteamientos paramétricos y no paramétricos para resolver esta cuestión en Schafer y Graham (2002) y Van Buuren (2018), para evitar la pérdida de registros o la imputación simple, que acarrea también inconvenientes (Rubin, 1976).

Los métodos tradicionales se basan en el criterio de ausencia de tipo aleatorio o MAR (“*missing at random*”), que se complementan con métodos de máxima verosimilitud y de imputación múltiple. Aunque existen métodos paramétricos que relajan la condición de la naturaleza aleatoria en la presencia de valores ausentes, como los basados en MNAR (“*missing not at random*”) (Van Buuren, 2018).

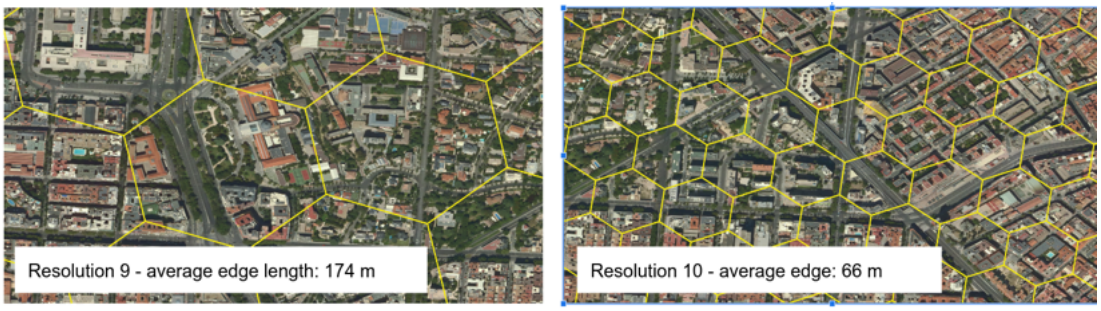
Además de los métodos paramétricos, existen métodos paramétricos que no asumen una distribución en la presencia de valores ausentes, como los métodos apoyados en los K-Vecinos más próximos (Fix y Hodges, 1989), o los basados en imputación en modelos de aprendizaje automático. Ambas familias permiten la imputación simple o múltiple de valores. La distinción entre imputación simple o múltiple procede de los modelos propuestos originalmente por Rubin (1976).

La primera usa un solo registro para imputar, mientras que la segunda utiliza un promedio de varios registros, y permite estimar la incertidumbre de la imputación.

La base de datos Idealista no cuenta de forma totalmente completa con el año de construcción, la altura de la finca, ni la superficie útil está siempre informada. Dado que estas variables son clave para el proceso de creación de los modelos de mercado, es necesario imputar estos valores cuando no existen, a través de imputación múltiple de tipo no paramétrico.

El primer proceso realizado es la imputación del año de construcción y la altura del edificio, consultando para ello la información catastral mediante un proceso de correspondencia espacial, en función de la finca en la que se ubica cada vivienda. Este proceso puede tener distintos grados de precisión, al utilizarse un método de imputación jerárquica sobre un índice espacial denominado H3 y desarrollado por la empresa Uber (2018). El proceso comienza intentando localizar la finca en un área pequeña (índice de resolución 13), y en el caso de no encontrarla iría a un área mayor (resolución 11), así sucesivamente hasta el área más amplia que sería la resolución 7. Una vez localizada la zona en la que se encuentra el inmueble se imputa la media del valor para ese área (año de construcción o altura).

Figura 2.17. Regiones de trabajo usando mallado hexagonal H3



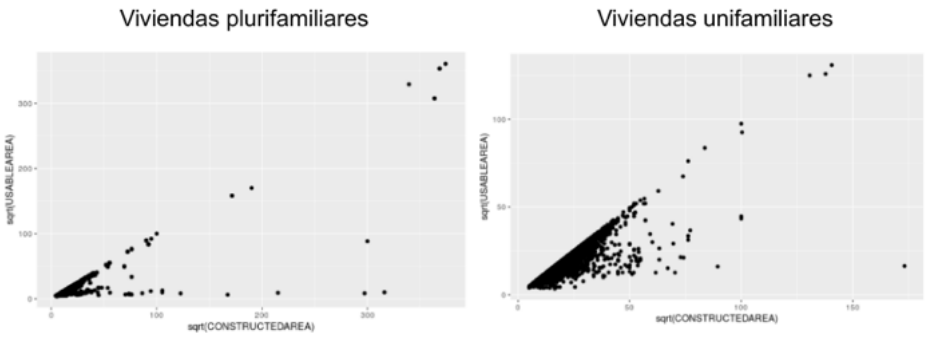
Fuente: elaboración propia.

El principio aplicado asume que en caso de no encontrar exactamente la finca en la que se localiza, se promedia el valor del área cercana, debido a que las zonas cercanas entre si están sujetas al mismo plan urbanístico y comparten características físicas. La Figura 2.17 muestra la rejilla construida por los índices H3 con resoluciones 9 y 10.

Otro atributo clave que se debe imputar es el área útil, que representa los metros totales que se pueden utilizar en una vivienda. En la base de datos Idealista sí se cuenta con el área construida, que guarda una fuerte correlación la útil, sin embargo, esta relación varía en función de la zona, el tipo de edificio y la rango de superficie. Para solventar esta cuestión, se ha decidido imputar el valor

mediante un árbol de regresión, utilizando el método CART[68] (Breiman, 2017), porque permite gestionar relaciones no lineales entre variables. Como se observa en la Figura 2.17 la proporción entre el área construida y el área útil guarda una relación lineal casi constante en el caso de los pisos, pero muestra un mayor grado de heterocedasticidad en las viviendas unifamiliares, de lo que se deduce que se necesita aplicar un modelo distinto para ese caso en concreto.

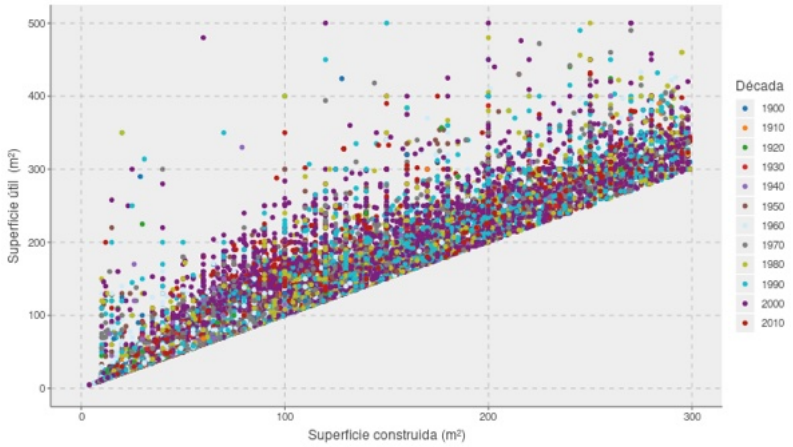
Figura 2.18. Relación entre área útil y área construida



Fuente: elaboración propia.

Se comprueba también que la relación entre el área construida y el área útil se mantiene estable en el tiempo (Figura 2.19).

Figura 2.19. Relación entre área útil y área construida, por década



Fuente: elaboración propia.

Finalmente se construyen dos modelos de árboles de regresión, uno para viviendas unifamiliares y otro para plurifamiliares. Se estima el cociente entre el área útil y la construida, y usa como variables independientes el año de construcción, el número de habitaciones y la zona geográfica en la que se ubica la vivienda. El modelo se expresa como sigue:

[68] En este caso se ha utilizado el paquete "rpart" de R (Therneau *et al.*, 2015).

$$R_{u,c} = \beta_0 + \beta_1 \cdot ANNOCON + \beta_2 \cdot NHABIT + \beta_3 \cdot LOCATION + \epsilon \quad [2.32]$$

donde $R_{u,c}$ representa el cociente entre la superficie útil y la construida, $ANNOCON$ el año de construcción, $NHABIT$ el número de habitaciones y $LOCATION$ una variable dicotómica para cada zona en la que se puede localizar el inmueble.

Anexo 2a. Propiedades axiomáticas de los números índices

Los índices de precios pueden estudiarse desde el punto de vista económico, en función de los conceptos con los que están relacionados (como el coste de vida o el precio de la vivienda), o bien desde el punto ángulo puramente matemático.

Desde el punto de vista analítico, Balk (1995) indica que la valoración formal de un índice óptimo se efectúa en función a una serie de propiedades que debe cumplir, que Diewert (2007) propone evaluar a través de nueve pruebas:

1. Prueba de identidad: si los precios se mantienen iguales y las cantidades se mantienen en la misma proporción con cada precio, entonces el índice tendrá un valor de uno. Cada cantidad de cada elemento se multiplica por el mismo factor α, para el primer periodo, o β para el periodo posterior.

$$I(p_{t_m}, p_{t_n}, \alpha \cdot q_{t_m}, \beta \cdot q_{t_n}) = 1 \forall\ (\alpha, \beta) \in (0, \infty)^2 \quad [2.33]$$

donde $I(P_{t_0}, P_{t_m}, Q_{t_0}, Q_{t_m})$ es un índice de precios para un momento del tiempo t, con un periodo base t_0; P_{t_0} y P_{t_m} son vectores que contienen los precios para desde t_0 a t; y Q_{t_0} y Q_{t_m} representan las cantidades para dichos periodos.

2. Prueba de proporcionalidad: si cada precio en el periodo original se incrementa por un factor α, entonces el índice se incrementará por el factor α.

$$I(p_{t_m}, \alpha \cdot p_{t_n}, q_{t_m}, q_{t_n}) = \alpha \cdot I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}) \quad [2.34]$$

3. Test de invarianza ante cambios de escala: el índice no debe cambiar si, en todos los periodos, los precios se incrementan por un factor y las cantidades se incrementan por otro factor.

$$I(\alpha \cdot p_{t_m}, \alpha \cdot p_{t_n}, \beta \cdot q_{t_m}, \gamma \cdot q_{t_n}) = I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}), \forall (\alpha, \beta, \gamma) \in (0, \infty) \quad [2.35]$$

4. Prueba de conmensurabilidad de las cantidades: el índice no debería estar afectado por el tipo de unidades seleccionadas para medir los precios o las cantidades.

5. Tratamiento simétrico del tiempo (o en paridad de medidas): revertir el orden de los periodos de tiempo debería ofrecer un número índice recíproco. Si el índice se calcula desde el periodo más reciente al más antiguo, este debería

ser el recíproco del índice calculado desde el periodo más antiguo al más reciente.

$$I(p_{t_n}, p_{t_m}, q_{t_n}, q_{t_m}) = \frac{1}{I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n})} \quad [2.36]$$

6. Simetría de las ponderaciones de cantidades: todas las cantidades deberían tener un efecto simétrico en el índice. Las permutaciones sobre el vector de componentes no deberían afectar al índice.

7. Prueba de monotonía: un precio posterior menor ($t+1$) que el precio en t, debería dar lugar a un índice menor que un índice de precios con un precio posterior mayor.

$$I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}) \leq I(p_{t_m}, p_{t_r}, q_{t_m}, q_{t_r}) \quad \Leftarrow \quad p_{t_n} \leq p_{t_r} \quad [2.37]$$

8. Prueba del valor medio: el precio general relativo debe estar entre el menor y mayor de los precios relativos para todas las cantidades.

9. Prueba de circularidad o transitividad: dados tres periodos ordenados t_m, t_n, t_r, la transitividad implica que una comparación directa entre las situaciones t_m y t_r dará el mismo resultado que una comparación indirecta entre t_m y t_r vía t_n. Aunque esta prueba fue propuesta por Fisher (1922a), siempre ha sido muy controvertida, hasta el punto que el propio Fisher la abandonó finalmente.

$$I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}) \cdot I(p_{t_n}, p_{t_r}, q_{t_n}, q_{t_r}) = I(p_{t_m}, p_{t_r}, q_{t_m}, q_{t_r}) \quad \Leftarrow \quad t_m \leq t_n \leq t_r \quad [2.38]$$

El proceso de selección del índice más adecuado debe perseguir el cumplimiento del mayor número de las nueve pruebas señaladas, garantizando así que la elección de este número índice proporcione una medida válida y confiable para el análisis económico subsecuente.

De este modo, se resalta la importancia del rigor metodológico en la implementación y uso de índices de precios, pues su validez no solo recae en su relación empírica con los fenómenos económicos, sino también en su adhesión a principios lógico-matemáticos.

Capítulo 3

Modelo de mercado

"Todos los modelos son erróneos, pero algunos son últiles."

— George E.P. Box

3.1 Introducción

El modelo de mercado tiene como objeto ser un estimador, con alto grado de detalle, de los precios de alquiler para el colectivo de estudio. Dado que se desconocen los datos individuales del mercado del alquiler, se construirá un modelo que calcule dicha información para los distintos estratos en los que se divide la población (geográficos y funcionales).

Se cuenta con tres fuentes de información de partida que deben relacionarse: los datos del Censo de Población y Viviendas, la Encuesta de Presupuestos Familiares y la muestra del portal inmobiliario Idealista. Las dos primeras contienen información agregada sobre la composición y precios del mercado del alquiler, y la tercera, cuenta con datos desagregados pero del mercado de oferta.

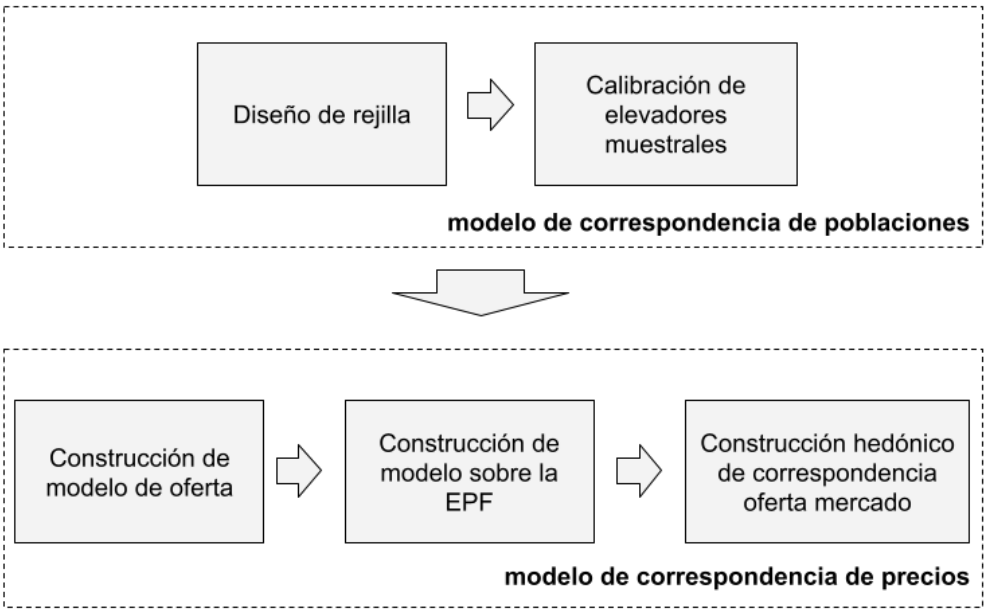
Por otra parte, es habitual que los registros de portales inmobiliarios cuenten con duplicidades o fenómenos de sobre e infrarrepresentación para ciertos segmentos del mercado (Loberto *et al.*, 2018; Pangallo y Loberto, 2018). Algunos de ellos, motivados por la diversidad de patrones de comportamiento de búsqueda en los distintos submercados inmobiliarios, tal y como presenta Boeing (2020) sobre datos de Craiglist en Estados Unidos. En su caso, la muestra de la plataforma de internet sobrerrepresentaba los estratos asociados a zonas donde había un uso más intenso de la tecnología, que se correspondían con áreas de mayor nivel económico.

El modelo a construir se basa en un proceso de correspondencia estadística

(*"matching estadístico"*) y persigue desarrollar un mecanismo capaz de relacionar los distintos conjuntos de información, aplicado a dos planos: el primero sobre la población, que calcula los elevadores muestrales del colectivo de oferta y el de alquiler; y el segundo, sobre los precios.

El proceso, descrito en la Figura 3.1, se compone de dos fases: el modelo de correspondencia de poblaciones y el modelo de correspondencia de precios de la vivienda.

Figura 3.1. Flujo de trabajo del modelo de mercado



Fuente: elaboración propia.

La primera fase se encarga de calcular los elevadores muestrales para cada registro Idealista, a través de un proceso de calibración de poblaciones en dos pasos. Los pesos estimados ofrecen una correspondencia entre el registro de oferta y un conjunto de hogares en régimen en alquiler, reduciendo la probabilidad de sesgos de sobrerrepresentación, infrarrepresentación y de no respuesta.

La segunda fase elabora un modelo hedónico anual que estima el precio de la renta, en euros/m²/año, para una vivienda a partir de una serie de características y un precio de oferta. Este modelo se denomina "modelo hedónico de correspondencia de precios", puesto que, calcula la relación entre los precios de las dos poblaciones.

El resto del capítulo se estructura en tres partes: la primera, describe los aspectos del diseño muestral, la segunda desarrolla la construcción de un modelo hedónico de mercado y la tercera, analiza en detalle los resultados obtenidos.

3.2 Modelo de correspondencia de poblaciones

La metodología planteada para el modelo de mercado consiste en el desarrollo de un proceso de correspondencia estadística (D'Orazio *et al.*, 2006). El modelo relaciona los datos de precios y la estructura poblacional de de la población de mercado a la de oferta.

La atribución de la estructura de la población se consigue mediante un proceso de calibración de los elevadores muestrales, mientras que los precios se resuelven a través de la imputación con modelos de precios hedónicos.

El modelo de correspondencia estadística poblacional, utiliza una estratificación basada en el grupo de variables comunes de los conjuntos de datos, que permite relacionar uno a uno los registros de oferta y mercado. Dado que se parte de información agregada y parcial de mercado[1], el proceso se realiza con el máximo nivel de desagregación posible de las fuentes de información involucradas.

3.2.1 Métodos de correspondencia estadística

Los métodos de correspondencia estadística (statistical matching)[2] para microdatos tienen como objetivo integrar dos o más fuentes de información ligadas a la misma población objetivo, para derivar una serie de datos sintéticos unificados en el que todas las variables estén disponibles de forma conjunta (D'Orazio *et al.*, 2006). El término "sintético" se refiere al hecho de que es una nueva base de datos construida mediante la imputación de las variables partir de los conjuntos de datos disponibles, a través de un estimador denominado de correspondencia, y no por medio de una unión directa de tablas mediante variables comunes conocidas.

Como indica Biancotti (2020), la creciente disponibilidad de nuevas fuentes de información permiten disponer de una visión más amplia y actualizada de la realidad, junto con la capacidad de mejorar o complementar el contenido de las estadísticas oficiales actuales, por ejemplo el Estudio piloto de movilidad del INE (2022d), basado en el posicionamiento de teléfonos móviles; o el uso de indicadores en tiempo real en la gestión de la crisis del Covid-19 en el Reino Unido (Rosenfeld, 2022). Estas nuevas fuentes conllevan dificultades de consistencia en la integración de los indicadores (Leucescu y Agafitei, 2013), puesto que, no es habitual disponer de fuentes de información con el mismo nivel de agregación o que permitan un cruce directo.

[1] Se dispone de una explotación de microdatos tanto para la EPF como para el Censo de Población y Viviendas, que no contienen datos de hogares sino datos agregados para estratos desglosados de hogares.

[2] Para más información consultar EUROSTAT (https://ec.europa.eu/eurostat/cros/content/statistical-matching-methods-method_en).

Por lo general, la correspondencia alinea las fuentes de datos comunes a través de atributos compartidos, o cuando existe, con otra información auxiliar[3]. En general, si la correspondencia se realiza sobre las variables compartidas por las fuentes de datos de partida, se asume el supuesto de independencia entre las variables no observadas (D'Orazio *et al.*, 2006) (independencia condicional).

Los conjuntos de datos sintéticos se pueden crear bajo tres enfoques: paramétricos, no paramétricos o mixtos. Todos ellos plantean dificultades metodológicas, tal y como recogen Leucescu y Agafitei (2013) en su revisión de metodologías de correspondencia estadística. Este trabajo, basado a su vez en el trabajo de D'Orazio (2006), desarrolla un estudio de viabilidad, metodológico y empírico, sobre la integración de varios conjuntos de microdatos de encuestas sociales en el marco de Eurostat. Su aproximación empírica une las encuestas europeas EU-SILC (condiciones de vida) y EQLS (calidad de vida), y además construye de un paquete de funciones estadísticas en lenguaje R denominado *"Statmatch"*[4].

El desarrollo de estas técnicas se inició la década de los 70s por Ruggles (1974), pero su adopción fue limitada debido a la imposibilidad de justificar y comprobar formalmente este tipo de relaciones (Kadane, 1978; Rodgers, 1984). Esta cuestión es una de las debilidades que muestran muchos de los métodos de correspondencia, particularmente los no paramétricos, ya que no es sencillo medir y comprobar la validez de las correspondencias (D'Orazio *et al.*, 2006; Rässler, 2012).

En general, estos procedimientos pueden resumirse como un método de imputación en la que existe un conjunto donante y otro receptor (Leucescu y Agafitei, 2013). El principio de general se puede explicar de la forma siguiente: si tomamos dos individuos i y j, de los conjuntos X e Y respectivamente y cuyos atributos son X_i e Y_j, si ambos individuos son suficientemente similares (según una serie de características comunes Z_i y Z_j), se podrían unir dando lugar a un nuevo registro sintético cuyos atributos fueran $X_i \cup Y_j$.

La condición de partida para estos procesos es que los conjuntos a vincular procedan de la misma población, aunque no exista una forma directa de relacionar las observaciones individuales. Esto los diferencia de un cruce de registros simple, donde se busca la correspondencia exacta por campos de unión conocidos.

La correspondencia estadística mide la relación entre registros en términos de similitud o distancia[5], de forma que dos registros se relacionarán si la similitud

[3] Por ejemplo una fuente de datos que contiene todas las variables interesantes o una estimación de una matriz de correlación, tabla de contingencia, etc.

[4] Para más información véase https://github.com/marcellodo/StatMatch.

[5] La distancia y la similitud son términos recíprocos, por tanto se utilizarán de forma indistinta.

entre ellos es suficientemente grande, o su distancia suficientemente pequeña.

El proceso tiene una dificultad adicional cuando no es posible disponer simultáneamente de los atributos del conjunto y los comunes, es decir, que no se observan conjuntamente los registros (X, Z) (Fuller, 2011). Por ello se proponen dos enfoques: el primero, se centra en las técnicas de análisis de incertidumbre (D'Orazio *et al.*, 2006; Rässler, 2012; Rubin, 1988) que trabaja macro-objetivos (estimación de una tabla de contingencia) en lugar de una tabla de microdatos; el segundo, busca la posibilidad de lograr la condición de independencia condicional usando información auxiliar. Para lograrlo, se puede utilizar: 1) conjuntos pequeños de subunidades con información completa sobre la distribución conjunta (Paass, 1986); o 2) variables *proxy* [6] con alto poder predictivo, para los casos donde la distribución conjunta de ciertas variables no es viable. Como indica Kott (2017), estas variables *proxy*[7] pueden mediar la relación entre Y y Z, y hacer plausible la condición de independencia condicional .

Todos los casos se centran en los estimadores de interés y no en la creación de conjuntos sintéticos (Schafer y Olsen, 1998). Por lo que los conjuntos generados no tienen por qué mantener los valores individuales originales, pero sí la distribución de los datos y la relación multivariante con las variables objetivo (Rubin, 1996). En consecuencia, es esencial controlar las dimensiones relevantes para el análisis y reflejar adecuadamente la incertidumbre asociada a los modelos implícitos.

Existen dos niveles la granularidad sobre los que aplicar la correspondencia: el macro y el micro. El primero, busca las relaciones entre variables no observadas de forma conjunta para estratos de la población, por ejemplo, la estimación de distribuciones conjuntas, marginales o matrices de correlación (D'Orazio *et al.*, 2006). Mientras que el enfoque micro, crea un fichero de microdatos sintéticos con todas las variables a partir de dos o más fuentes, relacionando los distintos conjuntos en función de sus variables compartidas.

El proceso de correspondencia asume el cumplimiento de una serie de requisitos previos, principalmente de armonización y coherencia. En este sentido D'Orazio (2006) propone 8 criterios a cumplir:

- Armonización en la definición de las unidades.
- Armonización del periodo de referencia.
- Completitud de la población.
- Armonización de variables.

[6] En estadística, una variable *proxy* es una medida que de forma individual es de poco interés, pero que permite obtener otras de mayor utilidad.

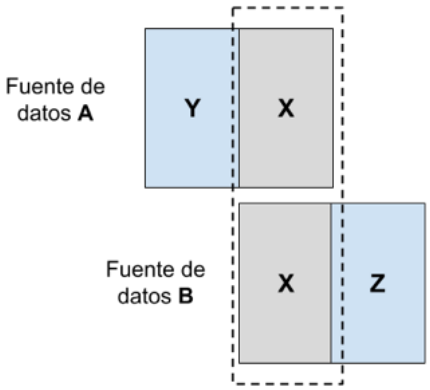
[7] En este caso se refiere a ellas como variables sombra (*shadow*) en lugar de "proxy".

- Armonización de clasificaciones.
- Ajuste por errores de medida (precisión).
- Ajuste por datos ausentes.
- Derivación de variables.

Una última cuestión a tener en cuenta es que aunque que existan atributos comunes, estos deben pertenecer a la misma población, por tanto, sus distribuciones tienen que ser muy similares. La similitud o discrepancia entre ellas se pueden medir con distintas medidas de distancia: la primera es la diferencia de la frecuencia ponderada de cada variable; la segunda, sobre la divergencia entre conjuntos, como por ejemplo, la de Helliger; y la tercera, basada en distancias entre distribuciones como la de Chi-cuadrado, Mahalanobis, la divergencia de Kullback-Leibler, o el test de Komogorov-Smirnov.

En el caso de que las distribuciones muestren diferencias sustanciales, se pueden utilizar los procedimientos de armonización, como la recategorización o calibración, para relacionar las distribuciones de los conjuntos donantes y receptores (Deville y Särndal, 1992).

Figura 3.2. Independencia condicional



Fuente: elaboración propia.

La elección de las variables comunes ejerce un enorme impacto en los resultados del proceso. Como apunta Adamek (1994), la selección de las variables adecuadas tiene más impacto que la técnica utilizada. La asunción de independencia condicional es el punto de partida del enfoque básico, y se puede describir gráficamente en la Figura 3.2. Es decir, si se dispone de dos conjuntos, A y B, referidos a la misma población, donde A cuenta con las variables X e Y, y el conjunto B dispone de las variables X y Z. Existe independencia condicional si las variables Z e Y son independientes, aunque la relación entre Z e Y pueda explicarse completamente por la variable Z (D'Orazio *et al.*, 2006). Lo anterior puede expresarse de forma funcional como:

$$f(x, y, z) = f(y|x) \cdot f(z|x) \cdot f(x) \quad [3.1]$$

donde X e Y son las variables del conjunto A, mientras que X y Z lo son del conjunto B, y f una función que toma como parámetro una o varias variables.

La independencia condicional es una propiedad esencial, ya que permite desarrollar una relación los dos conjuntos de variables A y B que garantiza que la distribución conjunta de las variables no comunes, Y y Z, sea la misma que la que se obtiene de un procedimiento de enlace perfecto. Dicha condición valida los procedimientos de asociación no observada y asegura que existe una fuerte relación predictiva entre las medidas donadas de un conjunto a otro.

Se pueden usar múltiples métodos para lograr un conjunto óptimo de predictores, como por ejemplo, la regresión *stepwise* o el análisis factorial. Otra forma de garantizar la validez del predictor es asegurando la calidad de las variables, por lo que es importante que estas no contengan errores, datos ausentes, ni tampoco, como apunta Scanu (2010), se recomienda usar variables altamente imputadas para hacer la unión.

Existe una gran diferencia entre las técnicas que requieren independencia condicional y las que no. Las primeras solo necesitan información en los conjuntos a unir, el resto de datos se usan solamente para comprobación. En las segundas, se puede usar información adicional para realizar la unión. En el caso de no asumir independencia, el tipo de enfoque dependerá de las características paramétricas del modelo. Si existe una distribución subyacente es posible usar técnicas paramétricas, sino, se utilizarían métodos no paramétricos. Existe un tercer enfoque que tiene que ver con el ámbito de aplicación, que atiende al nivel de agregación de los datos que se enlazan: micro, si son desagregados, o macro si son agregados.

3.2.1.1 Métodos y medidas de distancia

Leucescu y Agafitei (2013) afirman que los métodos más populares de correspondencia estadística son, con diferencia, los de tipo “micro” no paramétricos, bajo el supuesto de independencia condicional, y conocidos como imputación *“hot-deck”*. Estos métodos, se basan en imputar las variables no observadas en el fichero receptor con valores reales procedentes del fichero donante. Para medir la compatibilidad de la donación se utiliza una medida de distancia sobre las variables, de forma que la imputación de los valores sobre una observación en el fichero receptor, se realiza desde el registro más parecido dentro del fichero donante.

Existe un gran número de distancias básicas a aplicar como son la euclídea, manhattan o mahalanobis[8], o se puede utilizar una medida ponderada en función de la importancia de cada variable, a este método se lo conoce como distancia no restringida. Esta última puede dar lugar a que el mismo registro donante aporte información a varios registros receptores, lo que se denomina "poligamia", pero también es posible que ciertos registros del fichero origen se descarten al no haber correspondencia. Por otra parte, se pueden encontrar problemas con la distribución empírica de la variable Z imputada si no fuera idéntica a la distribución del fichero origen. Esta situación se puede evitar limitando el número de donaciones para cada registro origen.

Existe una alternativa restringida de donación que no permite que un registro donado solo se use más de una vez, y que es de especial utilidad cuando el fichero donante es mayor que el receptor. Consiste en minimizar la distancia entre los registros preservando la distribución de pesos en ambos conjuntos de datos, lo que asegura que la distribución empírica multivariante de las variables observadas se mantenga en el fichero sintético. Cuando hay más donantes que receptores se debe utilizar programación lineal, lo que exige una mayor carga computacional.

El enfoque paramétrico asume que la independencia condicional es suficiente para estimar los parámetros del modelo, por tanto, la función de verosimilitud conjunta se puede calcular como el producto de las verosimilitudes condicionales para cada conjunto de datos y la versosimilitud marginal de las variables comunes. En estos casos, se pueden emplear métodos de máxima verosimilitud (en adelante MLE), para estimar los parámetros de la distribución. Si bien es cierto, en ocasiones los estimadores de mínimos cuadrados se han empleado con resultados parecidos a MLE, como indica Rassler (2012). Aunque en todo caso, deben considerarse los inconvenientes asociados a los requisitos de los regresores ordinarios: la tendencia a la media de la regresión o que, en general, las especificaciones de los modelos suelen ser más imprecisas. Por estos motivos, los modelos de MLE son más atractivos en la práctica.

Existe un enfoque mixto que combina métodos paramétricos con los no paramétricos, intentando complementar la parsimonia[9] de los métodos paramétricos con la robustez y precisión de los no paramétricos. Un ejemplo de ellos es denominado *predictive mean matching imputation method,* propuesto por Rubin (1988), cuyo primer paso consiste en una regresión de los parámetros Z sobre X en la base de datos donante B. Dichos parámetros se usan entonces para estimar los valores de Z de la base receptora A. Finalmente, mediante una

[8] Para más información sobre las diferentes medidas de distancia, véase (Tan *et al.*, 2018).
[9] Según este principio ante dos métodos equivalentes, la mejor elección es aquel que es más simple.

función de distancia aplicada a cada registro a imputar con *hot-deck* en B, se decide si usar el valor paramétrico o el no paramétrico. Imputando aquel cuya distancia sea menor, de tal forma que se asegure una mayor probabilidad de que los valores finales mantengan la distribución de los datos originales.

Otro enfoque mixto es el basado en una puntuación de propensión (Rässler, 2012), expresada como la probabilidad condicionada de una unidad a pertenece a ambos grupos, dado un valor X. Por tanto, el valor a imputar será el registro cuya propensión sea igual o más cercana entre donante y receptor.

Los métodos anteriores se denominan simples, ya que toman una sola instancia para realizar la imputación. Una generalización de ellos son los denominados de imputación múltiple, basados en el trabajo de Rubin (1976), y que utilizan varios valores N para imputar cada valor ausente. El uso de varios registros permite mejorar la imputación, a la vez de que es posible estimar el grado la incertidumbre de la imputación[10].

Existen además métodos de correspondencia multivariante semiparamétricos basados en métodos bayesianos (Gómez-Rubio, 2020), que habitualmente se aplica a la imputación de valores ausentes. Estos modelos cuentan con la desventaja de requerir un alto esfuerzo de cálculo, pero a cambio, permiten medir el grado de incertidumbre de cada imputación.

3.2.2 Diseño de la muestra

El colectivo sobre el que se calcularán los índices de precios es el conjunto de viviendas en régimen de alquiler, para el periodo entre 2011 y 2019. De él, se dispone solamente de la estructura poblacional exacta en 2011, procedente de la explotación del Censo de Vivienda y Población del INE. Para los años 2012 y posteriores, se toma la información de la EPF con las rentas del alquiler pagadas por las familias, aunque el nivel de desglose de las mismas es mucho menor que el del censo. Además de las bases de datos anteriores, también se utilizan los datos de oferta del portal idealista, que cuenta con un nivel de desglose funcional y zonal muy profundo.

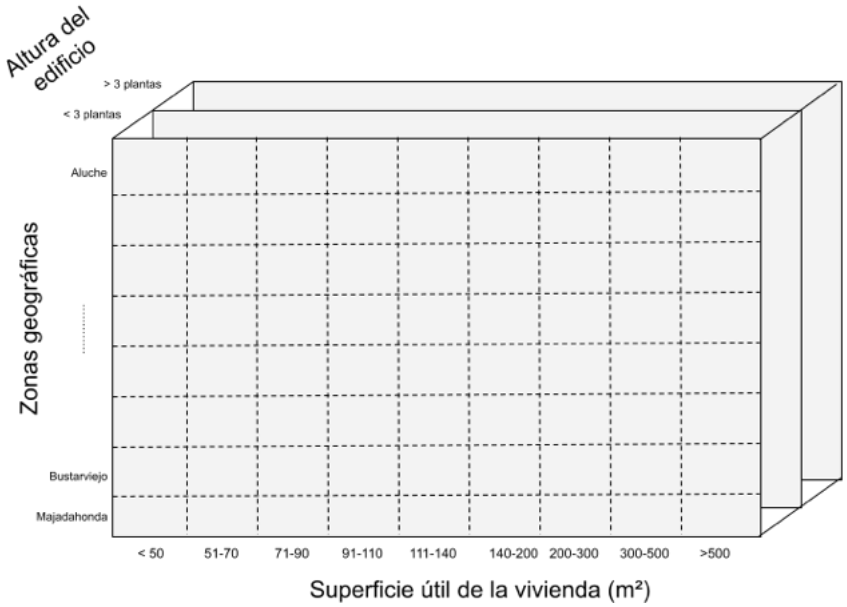
Para lograr una explotación detallada del índice de precios, se realizará una estratificación basada en el censo de viviendas que se aplicará sobre la base de datos de oferta. Como no se dispone del dato de censo a partir del segundo año y siguientes, se hará un ajuste de los pesos poblacionales del censo utilizando la evolución de la demanda, recogida por la EPF.

[10] Para más información sobre la cuestión de la imputación, véase el apartado 2.4.6.2 del capítulo anterior.

En este proceso se aplicarán tanto el término rejilla como las reglas de división o estratificación de la población, según las variables poblacionales de los registros de la muestra. Entre ellas, se encuentra la dimensión geográfica, cuyo máximo nivel de desglose es a nivel de barrios para Madrid capital, y municipios para el resto de la Comunidad de Madrid.

De cara a facilitar la comprensión de esta división, en la Figura 3.3 se presenta una versión simplificada de como se construiría una rejilla con los estratos de la muestra de viviendas. Para este ejemplo, se divide la muestra con tres variables de segmentación poblacional: rango de superficie útil, zona geográfica y altura del edificio.

Figura 3.3. Ejemplo de rejilla para una población de viviendas



Fuente: elaboración propia.

El proceso de creación de la rejilla, requiere la recodificación de las variables de las fuentes de oferta y censo para que los valores se expresen sobre las mismas escalas. Se calculan dos variables de totales sobre la rejilla: el número de viviendas y el total de superficie. Adicionalmente, será necesario imputar ciertos valores no observados que son necesarios para aplicar la reponderación[11].

3.2.2.1 Diseño de rejillas los procesos de calibración

Särndal y Lundström (2008) recomiendan que en la definición del vector de variables auxiliares se cumplan tres condiciones: (1) explicar bien el patrón de respuesta; (2) seleccionar correctamente las variables de estudio; y (3) identificar

[11] Véase apartado 2.4.6.2

los dominios de interés del análisis.

Sobre las condiciones anteriores, dada la diferencia en variables de las dos fuentes auxiliares para la calibración en el censo y la EPF, se define una rejilla para cada caso. Cada una de ellas se diseña sobre las variables comunes de la fuente con los datos de oferta. Las dos configuraciones de rejilla se describen en la Tabla 3.1, en la cual se observa que el número de dimensiones del la EPF es mucho más amplio, pero no identifica la zona geográfica exacta a la que pertenecen las observaciones. Por contra, en el caso del censo, no se incluye información del perfil sociodemográfico ni el gasto por hogar pero si la zona.

Tabla 3.1. Resumen de variables de cada rejilla de trabajo

Variable	**Descripción**	**Niveles distintos**	**Censo**	**EPF**
ANOCONSC	Año construcción	4	X	X
ASCENSOR	Ascensor	2	X	
GARAJE	Garaje	2	X	
LOCATION	Zona	178	X	
NHAB	Número Habitaciones	4	X	X
PLANTAS	Plantas	5	X	X
SUT	Superficie Útil	4	X	X
CAPROV	Código de provincia	1		X
DENSI	Densidad de población	3		X
FACTORGASTOT6	Factor de gasto	3		X
INTERINPSP	Ingresos netos	3		X
TAMAU	Población del municipio	5		X
TIPOCASA	Tipo de vivienda	3		X
TIPOEDIF	Tipo de edificio	4		X
ZONARES	Tipo de zona	3		X

Fuente: elaboración propia

Las variables utilizadas para la rejilla del Censo usan la misma nomenclatura que las variables de la fuente original (para más detalle véase epígrafe 2.4.2), y son las siguientes:

- *CAPROV*: Es capital de provincia o no.
- *LOCATION*: Código de Barrio en el caso de Madrid y Código de municipio (o grupos de municipios) para el resto.
- *SUT*: Superficie útil (4 niveles).
- *NHAB*: Número de habitaciones (4 niveles).

- *PLANTAS*: Número de plantas del edificio (5 niveles).
- *ASCENSOR*: Tiene ascensor (sí y no).
- *GARAJE*: Tiene garaje (sí y no).
- *ANOCONSC*: Año de construcción. Se usan 4 niveles, anteriores y hasta 1970, de 1971 a 1980, de 1981 a 1990, de 1991 a 2000 y de 2001 en adelante.

Para la calibración según la EPF se utiliza un número mayor de variables, que se enumeran a continuación[12]:

- *CAPROV*: Código de provincia (en este caso esta variable no es relevante porque solo se trabaja en el ámbito de la provincia de Madrid).
- *SUT*: Superficie útil (4 niveles), es importante tener en cuenta que este campo en la EFP viene limitado al intervalo entre 45 y 300 m^2.
- *NHAB*: Número de habitaciones (4 niveles).
- *ANOCONSC*: Año de construcción (4 niveles).
- *TIPOEDIF*: Tipo de edificio (4 niveles).
- *TIPOCASA*: Tipo de vivienda (3 niveles).
- *ZONARES*: Tipo de zona residencial (3 niveles). Se simplifica la clasificación original de la EPF en tres grados, zona de renta baja, media y alta.
- *INTERINPSP*: Intervalo de ingresos mensuales netos totales de cada miembro del hogar (ver definición en la descripción de la fuente de datos EPF). En este caso lo simplificamos a 3 niveles, renta baja (2 primeros niveles en la EPF), media (segundo y tercer nivel) o alta (tres niveles más altos de la variable en la EPF).
- *FACTORGASTOT6*: Gastos familiares, utilizando la codificación de la variable *factorGASTOT6* de la EPF. En este caso, se recodifican los valores originales de la EPF para tener 3 niveles: el primero, gasto bajo-medio que recoge los 4 primeros niveles iniciales; el segundo, gasto medio alto del quinto original; y el tercero, el gasto alto del sexto original.
- *DENSI*: Densidad de población del municipio (3 niveles).
- *TAMAMU*: Tamaño del municipio en población (5 niveles).

Dado que los datos de ingresos por miembro del hogar *INTERINPSP* y el factor de gasto *factorGASTOT6* son de utilidad para el enlace pero no se encuentran en el conjunto de oferta, se añaden en este último con un proceso de imputación múltiple no paramétrico[13] basado en modelos.

El modelo de imputación del factor de gasto se estima según las características de la vivienda y de la zona, a partir de los microdatos de la EPF para toda España. La forma funcional del modelo se define según la siguiente expresión analítica:

[12] El detalle de cada una de las variables de la EPF se describe en el epígrafe 2.4.1

[13] Se utiliza un modelo del tipo *Random Forests* mediante la librería *ranger* de R (Wright y Ziegler, 2015), para más información véase el Anexo 3b.

$$factorGASTOT6 \leftarrow TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + SUPERF + ANNOCON + DENSI + INTERINPSP + NHABIT + CCAA + CAPROV \quad [3.2]$$

donde $CCAA$ representa el código de comunidad autónoma. La configuración de este modelo utiliza 1.000 árboles, con un parámetro $mtry$[14] es igual a 10, y utiliza como pesos el factor de elevación de cada registro del fichero de la EPF.

Para el caso del modelo de ingresos, se utiliza el fichero de la renta media *per cápita* por sección censal de la Comunidad de Madrid. Al disponerse solamente de datos desde el 2016 en adelante, para los años 2015 y anteriores se asignan los ingresos del 2016.

3.2.2.2 Tratamiento de zonas geográficas

La segmentación zonal de la rejilla utilizada en la calibración censal trabaja con dos tipos de áreas[15]: la ciudad de Madrid y el resto de la provincia. En la ciudad, se toma como tamaño de área de trabajo el barrio, puesto que la sección censal no ofrece soporte suficiente de datos en oferta. Para el resto de la provincia se utiliza el municipio. En los casos en los que las celdas no cuentan con observaciones suficientes[16] se agrupan en zonas de orden superior según criterios de similitud inmobiliaria . El número total de zonas es 178, sumados barrios y áreas municipales .

Para la capital, los 128 barrios de la ciudad se agrupan finalmente en 112 zonas de trabajo, porque, como se ha indicado anteriormente, algunos no contaban con suficientes registros. Cada agrupación de dos barrios requiere cumplir el criterio de ser adyacentes geográficamente, y similares en características inmobiliarias y demográficas. Las zonas agregadas se denominan con el literal compuesto de los nombres de los barrios originales (ejemplo “El Pardo - Mirasierra”). Las zonas que se han unificado son:

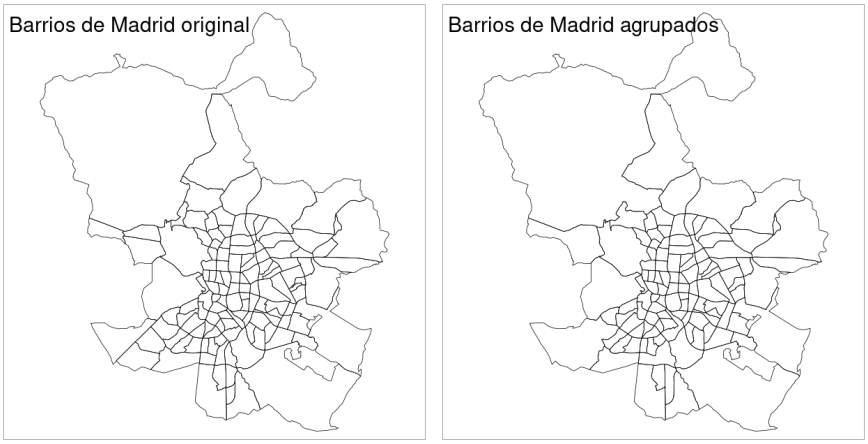
- Códigos 24 y 27: Legazpi - Atocha.
- Códigos 81 y 87: El Pardo - Mirasierra.
- Códigos 82 y 83: Fuentelarreina - Peña Grande.
- Códigos 104, 105, 106 y 107: Aluche - Campamento - Cuatro Vientos - Las Águilas.

[14] Indica el número de variables sobre las que se hacen los cortes del árbol de decisión.
[15] La selección de estos dos niveles está condicionada a las limitaciones de definición de detalle geográfico del Censo de Viviendas.
[16] Se ha tomado como requisito que cada celda geográfica debe contener al menos 30 registros por cada periodo anual.

- Códigos 121, 122 y 123: Orcasitas - Orcasur - San Fermín.
- Códigos 141 y 142: Pavones - Horcajo.
- Códigos 157 y 158: Colina - Atalaya.
- Códigos 161 y 162: Palomas - Piovera.
- Códigos 172 y 174: San Cristóbal - Los Rosales.
- Códigos 202 y 203: Hellín - Amposta.
- Códigos 211 y 212: Alameda de Osuna - Aeropuerto.

En la Figura 3.4, se muestran la división original y la final agrupada, se aprecia que no hay alteraciones sustanciales en la distribución geográfica.

Figura 3.4. Barrios de Madrid originales y agregados



Fuente: elaboración propia.

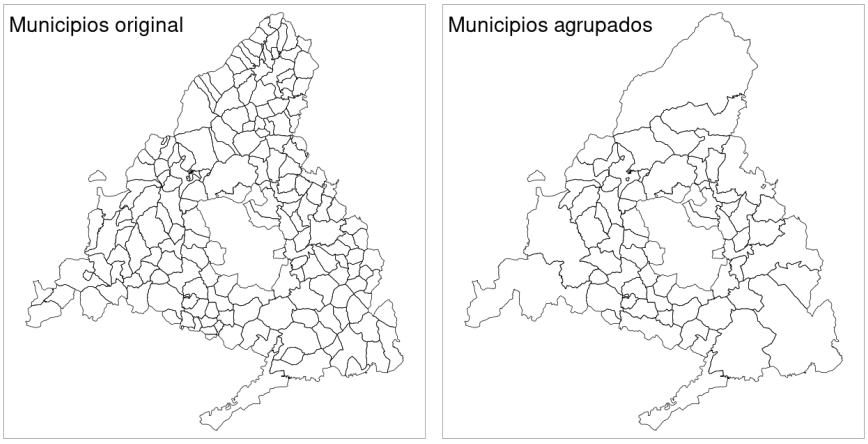
En el resto de la Comunidad de Madrid es más frecuente encontrar zonas de poco soporte, las cuales se han agrupado en municipios con un criterio parecido de similitud. Aunque en este caso, no se exige que el barrio sea adyacente sino que pertenezca a la misma área geográfica dentro de la Comunidad de Madrid (por ejemplo corredor del Henares, o municipios de la sierra de Guadarrama). Se crean las 22 nuevas zonas agrupadas:

- Área de Área de Manzanares el Real.
- Área de Área de Fuente el Saz de Jarama.
- Área de Área de la Sierra norte.
- Área de Daganzo de Arriba - Ajalvir.
- Área de la Cabrera - Torrelaguna.
- Área de Pedrezuela.
- Área de Guadalix de la Sierra.
- Guadarrama - Alpedrete.
- Área de Navacerrada.
- Área de Collado Mediano.

- Área de Mataelpino/Cerceda - Moralzarzal.
- Villanueva del Pardillo - Villanueva de la Cañada.
- Área de Brunete - Quijorna.
- Área de San Martín de Valdeiglesias - Cadalso de los Vidrios - Villa del Prado-Navas del Rey.
- Área de Colmenar del Arroyo.
- Área de Humanes.
- Griñón - Cubas de la Sagra - Casarrubuelos - Batres - Serranillos del Valle - Torrejón de Velasco.
- Área de Perales de Tajuña - Nuevo Baztán - Villarejo.
- Área de Colmenar de Oreja - Chinchón.
- Área de Villalbilla - Loeches.
- Área Meco.
- Área de Mejorada.

En este caso, como expresa la Figura 3.5, es necesario agrupar gran parte de los municipios rurales con menor población. Se parte de 178 municipios, que se consolidan en 66 zonas.

Figura 3.5. Municipios de la Comunidad de Madrid originales y agregados



Fuente: elaboración propia.

3.2.3 Reponderación de los elevadores muestrales

El proceso de correspondencia estadística que relaciona las poblaciones de oferta y la de mercado se realiza mediante un proceso de cálculo de elevadores muestrales de la oferta[17], por medio de un proceso de calibración. Dicho proceso de transformación de los pesos muestrales se denominará reponderación.

La reponderación es una técnica estadística usada comúnmente para compensar los errores de falta de respuesta y cobertura de una muestra. Como caso particular de ella, la calibración es un método que permite estimar los pesos adecuados para que la muestra reproduzca los totales de la población de estudio. Se utiliza habitualmente en muestreo de encuestas (Lohr, 2019), y mejora la calidad de la información mediante el uso de datos auxiliares, corrige de sesgos, y asegura el cumplimiento de suficiencia estadística de Fisher (Kullback, 2012) en la muestra de trabajo.

La falta de respuesta es uno de los principales problemas de cualquier proceso de muestreo, y es común encontrarla cuando se trabajan con datos de portales inmobiliarios (Bricongne *et al.*, 2023; Loberto *et al.*, 2018; Pangallo y Loberto, 2018). Lohr (2019) define el fenómeno de falta de respuesta unitaria como la ausencia de un registro completo, en nuestro caso, por ejemplo, la inexistencia de alquiler social en la muestra o la ausencia de registros de un tipo de vivienda en un estrato. Este fenómeno no debe confundirse con la presencia de valores no observados en los conjuntos de datos (como la ausencia del alguna variable en algunas de las observaciones). Los efectos de ignorar la falta de respuesta puede implicar serios problemas en los resultados del estudio (Fuller, 2011; Lohr, 2019).

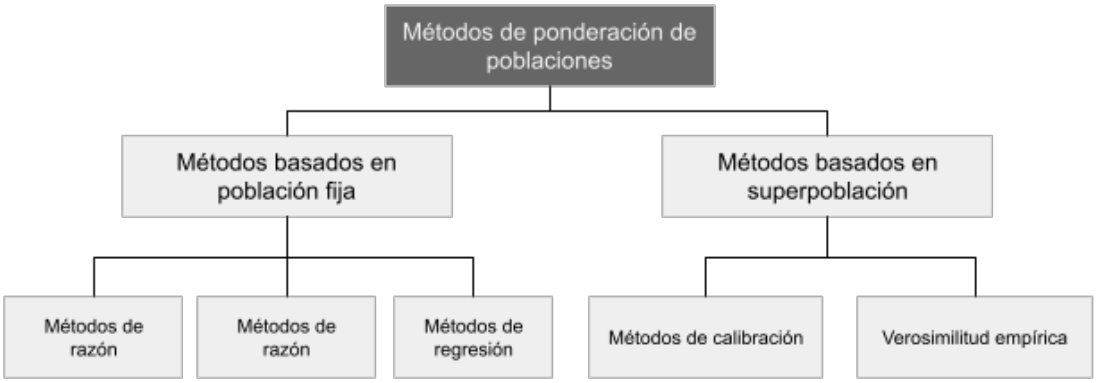
Lohr (2019) propone cuatro mecanismos para controlar la no respuesta: 1) prevención a través de un diseño muestral adecuado; 2) tomar una muestra representativa de los que no responden y usar esta submuestra para realizar una inferencia sobre el resto de los ausentes; 3) usar un modelo para predecir los casos del conjunto que no tiene respuesta; y 4) ignorarla, cuestión que se desaconseja totalmente .

Existen distintos enfoques para realizar la estimación de las ponderaciones poblacionales, resumidos en la Figura 3.6. Los primeros métodos propuestos sobre la base de una población fija, son los de estimación indirecta, entre los que destacan los métodos de razón, de diferencia y de regresión (Fuller, 2011). Todos ello se basan en modificar la forma original del estimador a través de la incorporación de las variables objetivo y auxiliares. Como estos métodos no se garantiza, de forma general, la reducción del error. Para solucionarlo, se

[17] Peso que indica la proporción de registros que representa una observación de la muestra.

propusieron posteriormente los modelos de superpoblación (Pérez-Villalta, 2002; Sánchez-Crespo, 2002) que generalizan los métodos de población fija.

Figura 3.6. Métodos de ponderación de poblaciones



Fuente: elaboración propia.

De forma más concreta, el método de muestreo en poblaciones finitas considera que los valores x_i, de la característica de interés asociados a una unidad u_i de una población finita U son fijos aunque desconocidos (excepto para los elementos de la muestra una vez que ha sido obtenida). Por tanto, esos valores no tienen la consideración de aleatorios. Dicha aleatoriedad, procede completamente de la selección de la muestra y se reflejan en el diseño muestral probabilístico. Los enfoques de población fija se basan en que los estimadores introducen variables indicadoras de la pertenencia de una unidad de la muestra, cuya distribución solo depende del sistema de selección usado.

Cassel (1977) indica que la aleatoriedad observada en una muestra puede tener de tres orígenes distintos:

- El método de selección de las unidades, que es el que se considera en el enfoque de clásico de las investigaciones por muestreo.
- Los métodos de medición de las variables en las unidades seleccionadas.
- El proceso que genera la verdadera medida de la variable para cada unidad, es decir que la fuente tiene cierta estructura aleatoria.

Los modelos de superpoblación, por otra parte, se apoyan principalmente en el uso de variables auxiliares X, cuyos valores son conocidos para todos los individuos de la población Y. Un dato x es un número real que procede de una población tras ser investigada o sujeta a experimentación. Si se repite la investigación o experimento, en general, se obtendría otro dato x', sobre el que si repetimos de en sucesivamente dará lugar a nuevos valores: x'', x''', y así sucesivamente. Estos valores pueden considerarse realizaciones muestrales de cierta variable aleatoria X. Este principio se denomina como muestreo repetido (Azzalini, 2017).

En consecuencia, si un dato es desconocido puede ser considerado una variable aleatoria, y cuando se conozca será una realización de ella. En este enfoque, el tipo de modelo más común es el de los estimadores de regresión, definidos por la expresión:

$$y_k = \mu(X_k) + \varepsilon_k \qquad [3.3]$$

donde ε_k es un error aleatorio, X_k la variable aleatoria e y_k la variable de interés, para la observación k, y μ la función de regresión que relaciona los predictores con la variable de interés.

La perspectiva de ponderación basada en modelos, propone que los estimadores de regresión generalizada son óptimos mientras que la población provenga de una superpoblación que siga un modelo de regresión lineal. Existen dos clases de estimadores basados en modelos: los estimadores de calibración (Deville y Särndal, 1992) y los de verosimilitud empírica (Chen y Qin, 1993; Chen y Sitter, 1999). Estos últimos, proponen un enfoque basado en la probabilidad empírica para el uso de información auxiliar. En nuestro caso, se usará un método de superpoblación basado en modelos a través de un estimador de calibración, que permitirá ajustar los pesos muestrales.

3.2.3.1 Calibración de los elevadores de oferta

La calibración es un método particular de correspondencia estadística que se aplica en estudios basados en encuestas, y que mejora la precisión de la estimación de parámetros sobre la muestra a través de información auxiliar. Es una técnica de uso habitual por las oficinas estadísticas nacionales europeas y norteamericanas.

La técnica se fundamenta en los procedimientos introducidos por Deming y Stephan (1940), y se propone formalmente por Deville y Särdnal (1992). En este trabajo se demuestra la equivalencia asintótica de la calibración[18] para el estimador de regresión generalizada (Cassel *et al.*, 1976), lo que garantiza las propiedades estadísticas de los estimadores calibrados. Posteriormente, Särndal (2007) propone una definición completa del método de calibración, que consiste en el cálculo de los pesos bajo unas restricciones, el cómputo de estimadores lineales ponderados sobre parámetros de la muestra y la construcción de un estimador casi insesgado (eliminando los sesgos de no respuesta y sobre e inframuestreo). Desde entonces su uso ha ido en aumento, y se han desarrollado nuevos métodos consecuencia de disponer computadores con mayor capacidad de cálculo. Entre ellos, los métodos basados en técnicas no paramétricas generales

[18] Para más información sobre el método de calibración, véase el Anexo 3a de este capítulo.

(Wu y Sitter, 2001); las basadas en regresiones locales polinómicas y las redes neuronales (Montanari y Ranalli, 2005); y optimizaciones en función de la medida de distancia[19] (Devaud y Tillé, 2019).

En la presente metodología, la calibración está orientado a ajustar las unidades de oferta y hacerlas corresponder con las unidades de mercado alquiler a lo largo del tiempo. Este proceso persigue la eliminación de sesgos de representación, de forma que todos los estratos de la muestra cumplan el criterio de suficiencia estadística de Fisher (Kullback, 2012). En su estado original, los registros de oferta tienen una serie de desequilibrios importantes, que son entre otros:

- Aún cuando el colectivo de oferta fuera igual al del mercado, el portal Idealista no representa al colectivo total de la oferta, a pesar de su amplia penetración de mercado; o bien que ciertos estratos socioeconómicos menos favorecidos estén menos representados en los portales inmobiliarios (Boeing y Waddell, 2017). Estas cuotas de mercado también varían a lo largo del tiempo (infrarrepresentación).
- El portal no cuenta con ciertos segmentos del alquiler, como alquiler social (falta de respuesta).
- En ciertas zonas, es habitual que el mismo inmueble esté anunciado por varias inmobiliarias (sobrerrepresentación) (Pangallo y Loberto, 2018; Wang *et al.*, 2020); o bien, los inmuebles que suscitan menos interés están sobrerrepresentados en la muestra de oferta (Han y Strange, 2016).

El dato de oferta cuenta con una extracción mensual de los inmuebles en el portal, mientras que las fuentes utilizadas para calibrar los pesos tiene frecuencia anual. Para utilizar la misma escala se agrupan las observaciones por código de anuncio[20] y año.

Las variables comunes X que se utilizan para establecer la correspondencia, deben existir tanto en el fichero Idealista como en la fuente estadística alternativa, y además es deseable que presenten una correlación lo más fuerte posible con las variables de interés Y. Para el caso del índice de precios, las magnitudes de interés serán los precios de oferta y de mercado.

Se realiza un proceso de reponderación en dos etapas (Lohr, 2019), que comienza con una primera calibración de la oferta sobre los totales del censo y una segunda calibración utilizando la información de la EPF. Esta segunda calibración se utiliza para hacer los ajustes temporales de los datos censales, por tanto solo aplica a los años 2012 y siguientes.

[19] Véase Anexo 3a de este capítulo.
[20] Variable Idealista *ADID* del fichero idealista.

Debido a que el nivel de profundidad de la información del que se dispone en el censo es distinto en la ciudad de Madrid que en el resto de municipios, se divide el proceso en dos calibraciones, uno para la ciudad y otro para el resto. Se describe el proceso en la Figura 3.7.

Figura 3.7. Métodos de ponderación de poblaciones

Fuente: elaboración propia.

Durante la construcción de los modelos de calibración se observan problemas de convergencia a partir de 2016, que se pueden la inestabilidad temporal del dato original de la EPF. Aunque no se puede determinar con exactitud la causa de la inestabilidad, ésta podría deberse del cambio metodológico en la EPF introducido a partir del año 2016 (INE, 2006a).

Para evitar los problemas de cambios abruptos en los pesos de la encuesta, se realiza un proceso de ajuste de los pesos poblacionales con suavizado exponencial sobre los totales de la EPF. El ajuste se realiza en función de su relación con los totales del Censo, de manera que, se intenta preservar el dato original pero asegurando una variación limitada con respecto al dato de referencia. La diferencia permitida entre ambas magnitudes es directamente proporcional a la diferencia en años entre 2011 y el año de la EPF que corresponda. De esta forma, se permiten variaciones mayores en años más lejanos al año base que en los años más cercanos.

El método de calibración seleccionado es una regresión logística con bandas[21] (logit). El estimador general de regresión, cuyo acrónimo es *GREG* (Deville y Särndal, 1992), se puede definir de forma coherente en muchas formas diferentes, por ejemplo: lineal, *raking*, truncado y *logit*. Se decide usar el logístico, o exponencial generalizado, porque permite controlar la aparición de valores extremos, es asintóticamente consistente, y ofrece siempre pesos positivos (Folsom y Singh, 2000).

El estimador de calibración generalizado calcula los pesos g, denominados *g-weights*, y calculados como $g_k = F(\lambda' z_k)$, donde z_k es un vector con valores definidos para registro $k \in s$ (o $k \in r$ donde r es el conjunto con respuesta

[21] Se ha utilizado el paquete *sampling* de R (Tillé y Matei, 2016).

conocida) y comparten la dimensión de un vector de variables auxiliares x_k. Los vectores z_k y x_k deben estar fuertemente correlados y el vector λ se determina a través de la ecuación de calibración:

$$\sum_{k \in s} d_k \cdot g_k \cdot x_k = \sum_{k \in U} x_k \qquad [3.4]$$

Si los vectores X_s y Z_s son iguales, se finaliza satisfactoriamente el proceso de calibración. En el caso del método *logit* los valores g, y por tanto los elevadores finales, estarán limitados entre un valor mínimo y máximo (Tillé y Matei, 2016). Los pesos g se pueden expresar también en función de la relación entre los pesos poblacionales finales y la distancia:

$$g_k = \left(\frac{w_k}{d_k} \right) \qquad [3.5]$$

donde w_k y d_k son el peso y la distancia para la observación k.

Dado que no existe un proceso para estimar los límites óptimos *a priori*, se desarrolla un proceso que se inicia con una banda superior e inferior amplias, y que va reduciendo progresivamente con un factor δ variable, por la parte superior e inferior, siguiensdo lo propuesto por Rao (1996). El proceso termina cuando no es posible reducir la parte superior o ampliar la parte inferior del intervalo.

Se utilizan varios niveles de precisión para encontrar el intervalo óptimo, que se muestran en la Tabla 3.2. El proceso comienza reduciendo el intervalo con una precisión baja hasta que es imposible reducir el valor. En ese caso, sobre el último valor válido, se intenta reducir el intervalo de forma sucesiva, hasta no poder mejorar el último intervalo válido.

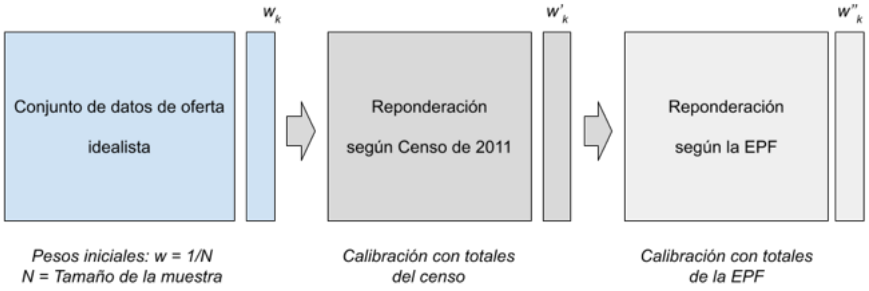
Tabla 3.2. Incrementos máximos y mínimos en bandas de calibración

Precisión	**Delta banda inferior**	**Delta banda superior**
Alta	0,002	-0,100
Media	0,005	-0,500
Baja	0,010	-1,000

Fuente: elaboración propia

En cada uno de los pasos del proceso se estimará un conjunto de pesos para todos los registros de la muestra de oferta, tal y como indica la Figura 3.8.

Figura 3.8. Cálculo de pesos



Fuente: elaboración propia.

Los pesos originales de la oferta w, se sustancian el elevador muestral w_k para cada registro k en el conjunto de datos. Este peso, se ajusta con la calibración en función de los totales del censo, para dar lugar a unos nuevos elevadores muestrales w'_k. Dado que estos pesos elevan la muestra al Censo de 2011, se aplica una nueva calibración para obtener los pesos definitivos de mercado w''_k, usando los totales de la EPF.

Los totales utilizados la calibración del censo han sido la superficie total útil y número de las viviendas en alquiler por estrato. Mientras que en la calibración de la EPF, se han utilizado el número de hogares en alquiler, a partir del factor de elevación de la muestra.

Las restricciones del modelo de calibración intentan mantener el equilibrio de la distribución de los totales y la estructura original de pesos de la oferta. En este sentido, se recuerda que en la calibración censal se controlan los totales por zona, mientras que en la segunda no, al no estar disponibles.

A modo de ejemplo, en la Tabla 3.3, se muestran los distintos pesos[22] que toman cinco anuncios a lo largo del proceso. Partiendo de los w_k originales de la tabla de anuncio se obtienen los w'_k, que son los pesos una vez reponderados según el censo, y finalmente los w''_k, una vez reponderados mediante la EPF. Para el primer anuncio se parte de un elevador muestral de 0,004%, es decir este anuncio representa ese porcentaje con respecto al total de viviendas, que teniendo en cuenta el censo representa a un 0,014% y vuelto a calibrar con la EPF llega a representar un 0,025% de las viviendas totales del colectivo.

[22] En este ejemplo, los pesos se expresan en porcentaje sobre el total poblacional.

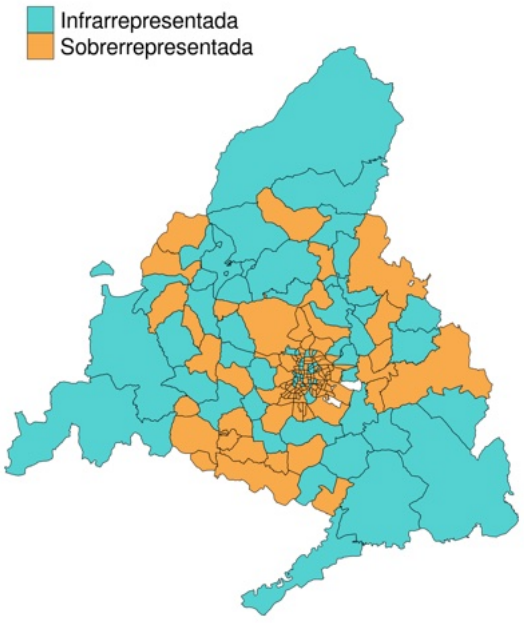
Tabla 3.3. Resultados de la reponderación, cambios en los elevadores muestrales en porcentaje

Código Anuncio	Peso Original	Peso Censo	Peso EPF
321578	0,019	0,069	0,260
2006331	0,019	0,043	0,221
2166585	0,030	0,098	0,198
25122953	0,019	0,044	0,216
25294841	0,023	0,107	0,206

Fuente: elaboración propia

En términos geográficos, la Figura 3.9 representa un mapa sobre el ajuste de los pesos en las distintas zonas para el año 2015. Cada zonas muestra si existe sobrerrepresentación o infrarrepresentación con respecto a su tamaño real. Se observa que principalmente las zonas centrales de la región tienden a la sobrerrepresentación, al contrario de las periféricas. Este fenómeno se podría atribuir al hecho de ser áreas con menor actividad inmobiliaria[23] y por tanto menor rotación de contratos.

Figura 3.9. Zonas sobrerrepresentadas e infrarrepresentadas en oferta



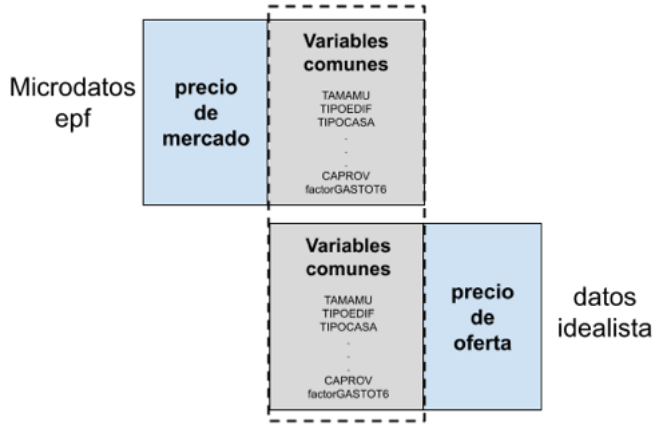
Fuente: elaboración propia.

[23] Según datos de idealista sobre el número de contactos medios por anuncios por zona.

3.3 Modelo de correspondencia de precios

El segundo método de correspondencia estadística es el de los precios, y relaciona los mismos en las bases de microdatos de Idealista y la EPF. Se construye mediante un modelo hedónico que tiene como covariables las características de la vivienda y el precio de oferta, y como variable objetivo, el precio del alquiler.

Figura 3.10. Modelo de correspondencia de precios



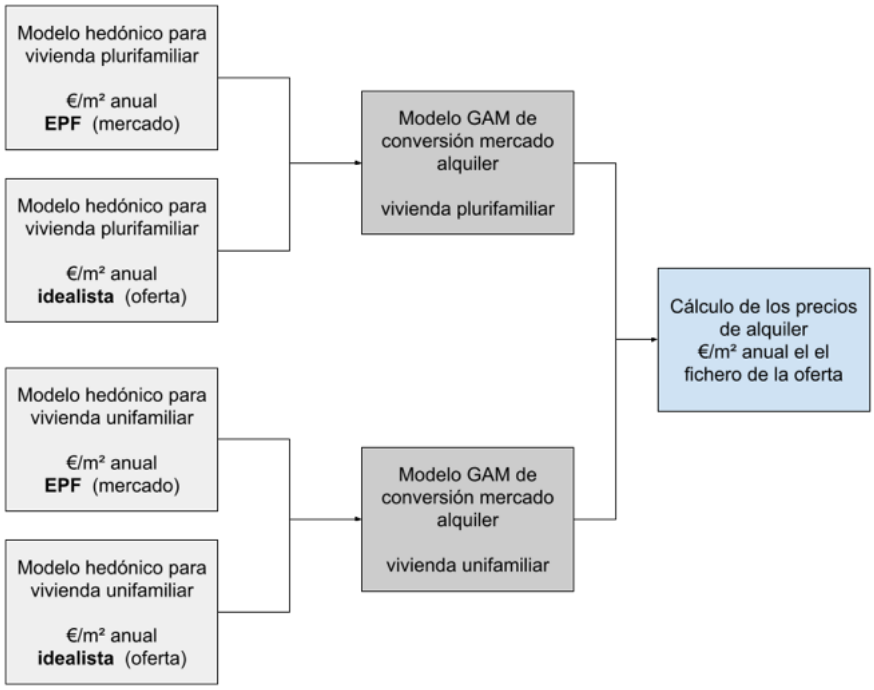
Fuente: elaboración propia.

Al desconocerse el precio de mercado de las viviendas a nivel individual para la oferta, primeramente, se realiza un proceso de imputación de precios de alquiler para todos los registros. Sobre estos datos, se construye un modelo lineal que traduce los precios individuales de oferta a su correspondiente precio de mercado. Este proceso de correspondencia estadística basada en modelos, se apoya en las variables comunes entre los datos de Idealista y la EPF, como se muestra en la Figura 3.10.

El modelo hedónico desarrollado se forma a partir de un conglomerado de 54 modelos (3 modelos x 2 tipos de vivienda x 9 años). El resultado final estima el precio anual del alquiler anual para cada registro del colectivo de estudio, que estará representado por el fichero Idealista con las ponderaciones de mercado.

El proceso completo se resume en la Figura 3.11, el cual se inicia con la creación de dos modelos de imputación de precios de mercado y oferta, que se unen con un último modelo que denominaremos de correspondencia de precios.

Figura 3.11. Descripción proceso de construcción de modelos hedónicos del mercado



Fuente: elaboración propia.

Siguiendo la recomendación de Eurostat (2014), se han desarrollado modelos hedónicos diferentes para cada tipología de vivienda residencial, unifamiliar y plurifamiliar. Ya que las diferencias en precios y características de ambos tipos son notables.

Para los 27 modelos hedónicos iniciales (oferta y alquiler), se han empleado modelos lineales aditivos generalizados[24] (en adelante GAM) (Hastie y Tibshirani, 2017), y árboles de regresión de tipo *Random Forests*[25] (Breiman, 2001). Se opta por los árboles por su mejor adaptación ante las características de la fuente de datos, que son son: la presencia de no linealidades, heterogeneidad espacial y heterocedasticidad (Baldominos *et al.*, 2018; Hastie *et al.*, 2017; Hjort *et al.*, 2022; Hong *et al.*, 2020; Valier, 2020). El Anexo 3c describe con detalle la técnica GAM, mientras que, el Anexo 3b detalla el método *Random Forests*.

El resumen de las técnicas aplicadas para cada hedónico se resume en la Tabla 3.4.

[24] Se ha utilizado el paquete *mgc*" de (Simon, 2017).
[25] Mediante el paquete "*ranger*" de R (Wright y Ziegler, 2015).

Tabla 3.4. Tipos de modelos creados

		Modelos		
Tipo	**Año**	**EPF**	**Idealista**	**Correspondencia**
Plurifamiliar	2011	R. Forests	R. Forests	GAM
	2012	R. Forests	R. Forests	GAM
	2013	R. Forests	R. Forests	GAM
	2014	R. Forests	R. Forests	GAM
	2015	R. Forests	R. Forests	GAM
	2016	R. Forests	R. Forests	GAM
	2017	R. Forests	R. Forests	GAM
	2018	R. Forests	R. Forests	GAM
	2019	R. Forests	R. Forests	GAM
Unifamiliar	2011	GAM	GAM	GAM
	2012	GAM	GAM	GAM
	2013	GAM	GAM	GAM
	2014	GAM	GAM	GAM
	2015	GAM	GAM	GAM
	2016	GAM	GAM	GAM
	2017	GAM	GAM	GAM
	2018	GAM	GAM	GAM
	2019	GAM	GAM	GAM

Fuente: elaboración propia

Para decidir qué método se utiliza en cada caso se ha aplicando el principio de parsimonia, de forma que, a igualdad de condiciones se prefieren lo modelos lineales. Sin embargo, en las viviendas plurifamiliares ha sido necesario utilizar *Random Forests* para lograr un correcto nivel de ajuste. Los motivos se pueden fundamentar en el hecho de que las viviendas unifamiliares representan el grupo más heterogéneo, en características y ámbitos geográficos, de la comunidad de Madrid.

El método *Random Forests* requiere el establecimiento de una serie de hiperparámetros, sobre las características de los árboles utilizados, las variables utilizadas en el proceso de construcción del modelo, el método para estimar la importancia de las variables y los pesos utilizados. Los pesos ponderan la importancia de las observaciones en la muestra, de forma que el método minimiza el error final del modelo ponderado según el peso poblacional de cada una de las

instancias. Los hiperparámetros aplicados se resumen en la Tabla 3.5.

Tabla 3.5. Hiperparámetros del modelos de mercado de tipo Random Forests

Modelo	Tipo	Número de árboles	Mtry	Tamaño mínimo nodo	Tipo importancia	Pesos
Mercado (EPF)	Plurifamiliar	1000	9	12	impurity	EPF

Fuente: elaboración propia

donde el número de árboles indica el número de estimadores que utiliza el algoritmo para calcular la estimación final; el tamaño mínimo de nodo es el número de observaciones mínimas asociadas a cada nodo hoja de los árboles; y el parámetro $mtry$ se refiere al número de variables sobre las que se puede aplicar una regla de decisión al construir los árboles. El tipo de importancia recoge al criterio de reducción de entropía[26] utilizado para establecer los cortes.

3.3.1 Modelos hedónicos básicos de mercado y oferta

Por cada tipo de vivienda se desarrolla un par de modelos hedónicos sobre un conjunto de variables comunes de las fuentes, Idealista y EPF. En el primer caso, se estima el precio del alquiler a precios de mercado, utilizando el conjunto de microdatos de la EPF[27]. La variable a predecir es el logaritmo del precio del alquiler anual por metro cuadrado útil, siendo su forma funcional la siguiente:

$$\begin{aligned} log(\hat{P}_m) \leftarrow & TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + \\ & SUPERF + ANNOCON + DENSI + INTERINPSP + \\ & NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.6]$$

donde $\hat{P}_m$ es el precio de alquiler anual de mercado por unidad de superficie útil, $TAMAMU$ el tamaño del municipio, $TIPOEDIF$ el tipo de edificio, $TIPOCASA$ el tipo de vivienda, $ZONARES$ el tipo de zona residencial, $SUPERF$ la superficie útil de la vivienda en m², $ANNOCON$ el año de construcción, $DENSI$ la densidad de población de la zona, $INTERINPSP$ los ingresos del cabeza de familia, $CCAA$ la comunidad autónoma, $factorGASTOT6$ los ingresos del cabeza de familia, $CAPROV$ es una variable dicotómica que indica si la observación está en la capital de provincia o no.

[26] En este caso la entropía se interpreta como el nivel de desorden en la variable objetivo, medido como índice Gini, que reduce el árbol al dividir la población con esta variable.
[27] Ver sección 2.4.1.

Posteriormente, se construye un modelo que estima la misma magnitud pero para la oferta ($\hat{P}_o$), en cuyo caso la fuente es el fichero Idealista sobre las mismas variables del modelo anterior, como se aprecia en su forma funcional:

$$\begin{aligned} log(\hat{P}_o) \leftarrow TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + \\ SUPERF + ANNOCON + DENSI + INTERINPSP + \\ NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.7]$$

3.3.2 Modelos de correspondencia de precios

Los dos conjuntos de precios, de mercado ($\hat{P}_m$) y oferta ($\hat{P}_o$), deben relacionarse a través de un modelo de correspondencia. El enlace se realiza también mediante un modelo, cuyo conjunto de variables comunes son las covariables presentes en ambos modelos más el precio de oferta. Al existir relaciones no lineales entre el precio de mercado y las covariables, se ha decidido utilizar un modelo lineal GAM, que ofrece un buen balance entre interpretabilidad y ajuste (Hastie y Tibshirani, 2017), y evita algunos inconvenientes de los modelos de aprendizaje estadístico aplicados al precio de la vivienda (Nguyen y Cripps, 2001), como el la selección correcta de hiperparámetros.

Las funciones de suavizado del modelo GAM permiten adaptar las contribuciones de las covariables a sus diferentes valores. El modelo de correspondencia de precios se especifica de la forma siguiente:

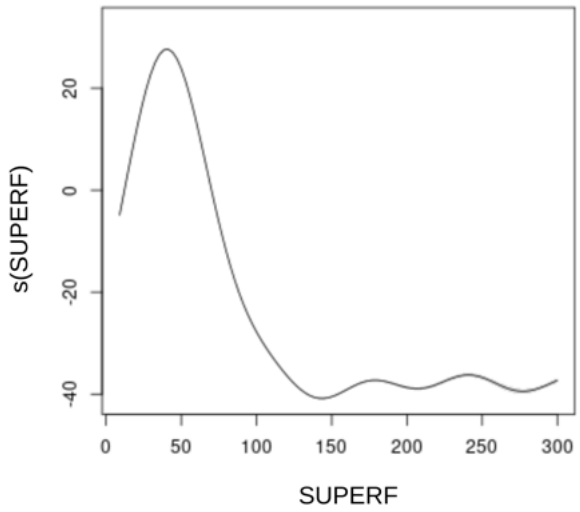
$$\begin{aligned} log(\hat{P}_m) \leftarrow s(\hat{P}_o) + TAMAMU + TIPOEDIF + TIPOCASA + \\ ZONARES + s(SUPERF) + ANNOCON + DENSI + INTERINPSP + \\ NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.8]$$

donde el término $s(SUPERF)$ indica que se aplica una función de suavizado, o *spline*, para modelar la relación entre la superficie y el precio. Las funciones de suavizado s se calculan mediante la agregación de una serie de funciones base (Hastie *et al.*, 2017), y son una generalización de los modelos lineales. De forma que una regresión por mínimos cuadrados usa una función de suavizado s en lugar de un coeficiente fijo, en cuyo último caso representarse geométricamente como una recta con pendiente constante (Hastie y Tibshirani, 2017) que representa la contribución del parámetro, en lugar de una curva en el caso de la mencionada s.

La Figura 3.12 muestra los valores de la suavizado s para el la variable *SUPERF*

en el modelo de correspondencia oferta-mercado. La función se puede interpretar como un coeficiente dinámico para los distintos valores de la variable. En la Figura se observa como para superficies pequeñas el coeficiente es positivo hasta llegar a 50 m² aproximadamente. Entre 50 m² y 80 m², la contribución es positiva pero decreciente, mientras que a partir de 80 m², la contribución al precio es negativa. Para valores superiores la relación es decreciente, hasta que los 140 m², donde los incrementos en la superficie no implican incrementos significativos en el precio.

Figura 3.12. Valores de la función de suavizado (s) para la variable superficie útil en modelo de correspondencia



Fuente: elaboración propia.

El modelo de correspondencia permite estimar el precio de mercado para cualquier registro de la oferta. Por tanto, dado el alto nivel de detalle de los atributos del conjunto de datos de la oferta y los elevadores muestrales, es posible construir series de precios de mercado altamente desagregadas. Sin embargo, el conjunto de atributos utilizados para estimar los precios de oferta no es muy exhaustivo, por lo que el modelo adolece sesgos por omisión de variable. Para solucionar esta cuestión y lograr un mayor nivel de precisión en las series de precios, será necesario ajustar el resultado del modelo de correspondencia con un modelo hedónico más preciso (que es el hedónico de oferta que se presentará en el Capítulo 5).

Por otra parte, el método de correspondencia no utiliza como variables de áreas geográficas, ya que los microdatos de la EPF no disponen de dicha información. Debido a que las cada zona tiene un comportamiento particular, será necesario un proceso posterior de control para asegurar el correcto control de la heterogeneidad espacial sobre los resultados.

3.4 Resultados

El proceso de correspondencia asigna, para cada registro de la oferta, un elevador muestral y un precio de mercado. Para evaluar la calidad de los resultados se estudiará el cumplimiento de las condiciones propuestas por Rässler (2012):

- Preservar los valores individuales: se comprueba que se respetan las sumas de los pesos totales para los distintos criterios de estratificación, además de asegurar que las diferencias entre el peso original y el final se encuentran dentro de bandas aceptables.
- Preservar la estructura de la correlación: se analiza si el precio de mercado estimado mantiene un buen nivel de ajuste con respecto a sus valores originales. Para ello, se estudiará la calidad del modelo de correspondencia de precios.
- Preservar la distribución conjunta: se valida si la distribución conjunta de los pesos muestrales finales no difiere, de forma sensible, con respecto a la distribución conocida del fichero de la EPF.
- Preservar las distribuciones marginales: analiza que las distribuciones marginales de variable objetivo se corresponden con las originales. De forma particular, se estudia el efecto en la distribución espacial del modelo, dado que, el proceso de reponderación utiliza parcialmente esta información para la estimación de los pesos.

3.4.1 Valores individuales

Los pesos individuales de la calibración deben cumplir los requisitos de ajuste, es decir, ser positivos y estar acotados entre un valor máximo y mínimo. Al cumplirse la condición de convergencia en el proceso de calibración, expresada como el máximo valor de la diferencia entre los totales, se asegura que la suma de los totales de cada una de las variables no presentan diferencias apreciables. El criterio de convergencia asegura la siguiente condición:

$$\frac{max(X_s \cdot g_s \cdot d_s - T_s)}{T_s} \leq 10^{-6},\ \forall s \in S \qquad [3.9]$$

donde T_s representa el total para cualquier estrato s de los S criterios de estratificación; X_s son los valores individuales de las variables; y $g_s \cdot d_s$ los nuevos pesos muestrales. La condición anterior asegura que el ratio entre la diferencia de los totales ponderados ($X_s \cdot g_s \cdot d_s$) con respecto a los totales T_s es mayor que uno entre un millón.

La condición anterior se cumple para los límites máximo y mínimo de las g

distancias en cada uno de los procesos de calibración, calculados de forma iterativa (Rao, 1996). La Tabla 3.6 contiene las bandas finales de calibración[28] para el Censo, siendo el rango más amplio de 0,05, para la parte inferior, a 10 para la superior. Se observa que el rango se amplia, especialmente para los años 2018 y 2019, pero incluso en este caso no representan *g-distancias* extremas (D'Orazio *et al.*, 2006). En términos zonales, no se aprecia diferencias importantes entre Madrid y el resto de la provincia.

El factor de 10 en 2019, significa que un registro de oferta representaría 10 hogares en alquiler. Para un factor de 0,05, en la banda inferior, se necesitarían 20 registros de oferta para representar un registro del mercado.

Tabla 3.6. Bandas de calibración Censo y EPF

	Censo				EPF	
	Madrid		Resto CAM		Todas las zonas	
Año	**Inferior**	**Superior**	**Inferior**	**Superior**	**Inferior**	**Superior**
2011	0,28	4,50	0,27	6,67		
2012	0,24	6,00	0,24	6,95	0,46	1,90
2013	0,20	5,00	0,20	5,00	0,58	2,00
2014	0,25	4,00	0,20	3,00	0,28	3,00
2015	0,26	4,00	0,16	4,50	0,38	4,00
2016	0,18	3,00	0,18	3,25	0,38	3,00
2017	0,10	3,50	0,10	4,50	0,08	5,00
2018	0,05	10,00	0,05	10,00	0,10	5,50
2019	0,05	10,00	0,05	10,00	0,10	5,50

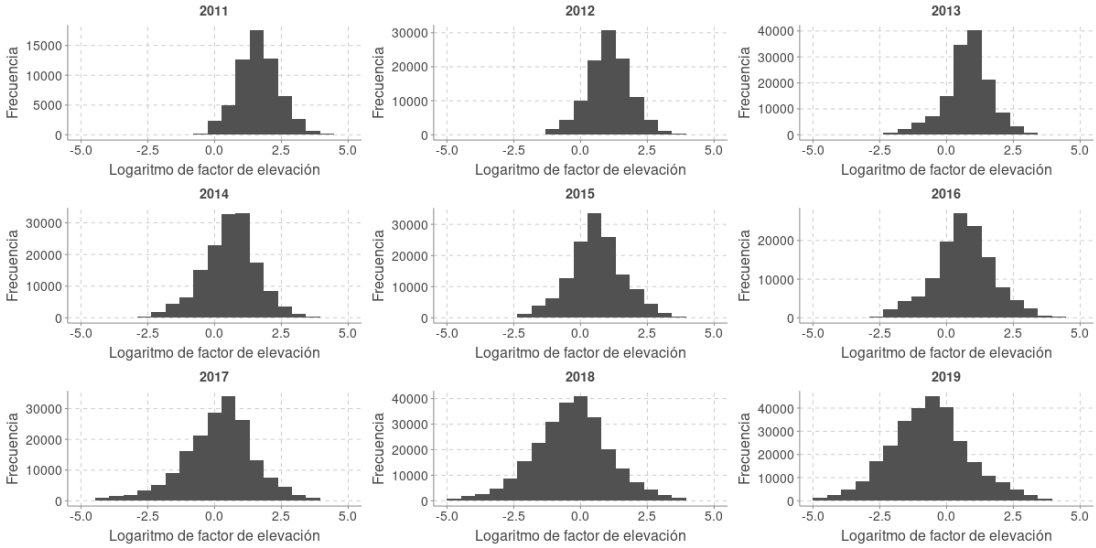
Fuente: elaboración propia

En la calibración de la EPF las bandas son más estrechas que en el caso del Censo. Sin embargo, como sucede en el caso anterior, los intervalos se van ampliando en los últimos años de la serie, lo cual es lógico al estar más alejados del periodo base.

Como se ha indicado anteriormente, la EPF muestra cierta inconsistencia temporal a partir del 2016 que no se corresponden a cambios en el mercado, y que por tanto se puede atribuir a las modificaciones metodológicas introducidas a partir de ese año. En el epígrafe del Anexo 3d, se adjuntan la Tabla 3.17 de totales originales, y la Tabla de 3.18 totales suavizados , en ellas confirma un notable cambio de escala en el año 2016. Los datos una vez ajustados reducen esta discontinuidad, por ejemplo, la variación del total poblacional original entre 2015 y 2016 es del 12%, mientras que el suavizado la reduce al 2%.

[28] Valores mínimos y máximos del factor $g \cdot d$ que ajusta los pesos originales.

Figura 3.13. Distribución de los factores de elevación en escala logarítmica



Fuente: elaboración propia.

En cuanto a la distribución de los pesos muestrales a lo largo del tiempo, se observa que el rango de los factores de elevación finales tiende a ampliarse en los últimos años de la serie (véase Figura 3.13). Además, se observa como los elevadores inferiores a 1 son prácticamente inexistentes en 2011, mientras que en 2019 son mayoritarios, representando más de la mitad de los casos.

Los cambios en la forma de la distribución pueden tener diversos orígenes. Uno específico por la composición de los anuncios publicados en el portal, que de forma progresiva va ampliando su penetración en el mercado. Por otra parte, al tratarse de un mercado en expansión, la relación entre oferta y mercado varía en el tiempo (Ardila *et al.*, 2021; De Wit *et al.*, 2013; Han y Strange, 2016). Ambos motivos dan lugar a que la relación *stock* en alquiler / *stock* en oferta sea progresivamente menor, produciendo elevadores muestrales decrecientes, tal y como vemos en la Figura 3.13, donde el centro de masa de las distribuciones pasa de valores positivos en 2011 a valores ligeramente negativos en 2019.

3.4.1.1 Implicaciones de la estabilidad temporal de la muestra

Con el objetivo de comprobar que la estratificación de origen es comparable a lo largo del tiempo[29], se ha realizado un análisis preliminar para evaluar si la estructura de estratos se mantiene estable a lo largo del tiempo. Para ello, se han tomado las poblaciones de Tres Cantos y Fuentelsaz del Jarama, por representar realidades inmobiliarias distintas. La primera, es una población residencial de la zona metropolitana con un nivel de ingresos medio-alto; y la segunda, una zona

[29] Debido a los cambios metodológicos aplicados sobre la EPF en el periodo de análisis.

rural. En la Tabla 3.7 y la Tabla 3.8 se muestra la presencia de los estratos a lo largo del tiempo (con una “X” si está presente en la muestra y vacío cuando está ausente). En ambos municipios se observan cambios de composición a partir del año 2015 y 2016, que, como se comentaba anteriormente, es el momento en el que la EPF aplica una nueva metodología de trabajo.

Para el caso de Fuentelsaz, solo se dispone de información para el segmento de densidad de población intermedia hasta 2013, mientras que, la zona diseminada mantiene una estructura estable a lo largo del tiempo. Este comportamiento puede ser atribuible a que su población es rural y el tipo diseminado es el mayoritario, por tanto, existe soporte de datos en todos los periodos.

En Tres Cantos, sin embargo, se produce un cambio en la estructura de los estratos que representan el municipio entre 2016 y 2017. Antes de 2017, la muestra se encuentra en zonas densamente pobladas del municipio, mientras posteriormente pasa a concentrarse en zonas con densidad intermedia. Esto en términos intramunicipales no debería tener un efecto importante, pero si lo podría tener en el cálculo de los totales agregados, por densidad de población, en la Comunidad de Madrid.

Este análisis se ha realizado únicamente en los municipios de la Comunidad de Madrid, a excepción de la capital, por cuanto la muestra en la última se mantiene estable.

Tabla 3.7. Presencia de estratos en Fuentelsaz del Jarama (todos excepto casa económica)

Densidad	Tipo	Edificio	Hab.	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
diseminada	Casa media	10 ó más	1 o 2	Urbana media									X
diseminada	Casa media	menos de 10	1 o 2	Urbana alta						X			
diseminada	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Casa media	10 ó más	3	Urbana media			X	X	X	X	X	X	
diseminada	Casa media	menos de 10	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Casa media	menos de 10	4	Urbana media		X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	1 o 2	Urbana media					X	X	X	X	
diseminada	Chalé o casa grande	independiente	1 o 2	Urbana alta						X			
diseminada	Chalé o casa grande	independiente	1 o 2	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	4	Urbana alta						X			
diseminada	Chalé o casa grande	adosada	4	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	4	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	5 o más	Urbana alta						X			
diseminada	Chalé o casa grande	adosada	5 o más	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	5 o más	Urbana media			X					X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X					
intermedia	Casa media	10 ó más	3	Urbana media			X	X					
intermedia	Casa media	menos de 10	3	Urbana media	X	X	X	X					
intermedia	Casa media	menos de 10	4	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	1 o 2	Urbana media	X		X	X					
intermedia	Chalé o casa grande	independiente	1 o 2	Urbana media	X	X							
intermedia	Chalé o casa grande	adosada	3	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	independiente	3	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	4	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	5 o más	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	independiente	5 o más	Urbana media				X					

Fuente: elaboración propia

Tabla 3.8. Presencia de estratos, municipio de Tres Cantos (casa de tipo medio)

Densidad	Tipo	Edificio	Hab.	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
densa	Casa media	10 ó más	1 o 2	Urbana alta	X	X	X	X	X	X			
densa	Casa media	10 ó más	1 o 2	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	1 o 2	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	3	Urbana alta	X	X	X	X	X	X			
densa	Casa media	10 ó más	3	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	3	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	3	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	4	Urbana alta		X	X	X	X	X			
densa	Casa media	10 ó más	4	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	4	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	4	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	5 o más	Urbana media				X	X	X			
densa	Casa media	menos de 10	5 o más	Urbana alta				X					
densa	Casa media	menos de 10	5 o más	Urbana media		X	X						
intermedia	Casa media	10 ó más	1 o 2	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	1 o 2	Urbana media							X	X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana alta							X	X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana media							X	X	X
intermedia	Casa media	10 ó más	3	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	3	Urbana media							X	X	X
intermedia	Casa media	menos de 10	3	Urbana alta							X		X
intermedia	Casa media	menos de 10	3	Urbana media							X	X	X
intermedia	Casa media	10 ó más	4	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	4	Urbana media							X	X	X
intermedia	Casa media	menos de 10	4	Urbana alta							X	X	X
intermedia	Casa media	menos de 10	4	Urbana media							X	X	X
intermedia	Casa media	10 ó más	5 o más	Urbana media							X	X	X
intermedia	Casa media	menos de 10	5 o más	Urbana media									X

Fuente: elaboración propia

3.4.2 Estructura de la correlación

Para evaluar que la estructura de correlación entre la muestra y los resultados del modelo se mantiene constante, se evaluará la calidad del ajuste de los modelos construidos[30], medida en R^2 ajustado. Se toma esta métrica porque es aplicable tanto a modelos lineales como a árboles de regresión, y es más informativa que otras medidas, como el error cuadrático medio o el error medio absoluto (Chicco *et al.*, 2021).

El R^2 es una medida estadística que recoge la proporción de la varianza de la variable dependiente explicada por un modelo sobre las variables independientes. Toma valores entre $-\infty$ y 1, donde 1 es el ajuste perfecto de un modelo capaz de capturar cualquier comportamiento de la variable objetivo. Mientras que 0 indica que el modelo no puede capturar ninguna relación entre los regresores y la variable objetivo, y los valores inferiores a cero muestran un modelo que ajusta peor que una linea horizontal (Chicco *et al.*, 2021).

De forma general, el R^2 se puede calcular según la expresión:

$$R^2 = 1 - \frac{\sigma^2_{error}}{\sigma^2_y} \qquad [3.10]$$

donde σ^2_{error} es la varianza de los errores del modelo, y σ^2_y es la varianza de la variable dependiente y.

En nuestro caso se opta por una versión ajustada del R^2, que soluciona dos problemas importantes de la medida original: el primero, reduce el sobreajuste asociado un número elevado de grados de libertad del modelo; y el segundo, la propensión del R^2 a ofrecer valores más altos cuanto mayor es el número de parámetros del modelo. Intenta, además, que la magnitud exprese el porcentaje de la variable explicado solamente por los regresores que afectan a la variable dependiente. Por tanto, la versión ajustada del coeficiente de determinación sería:

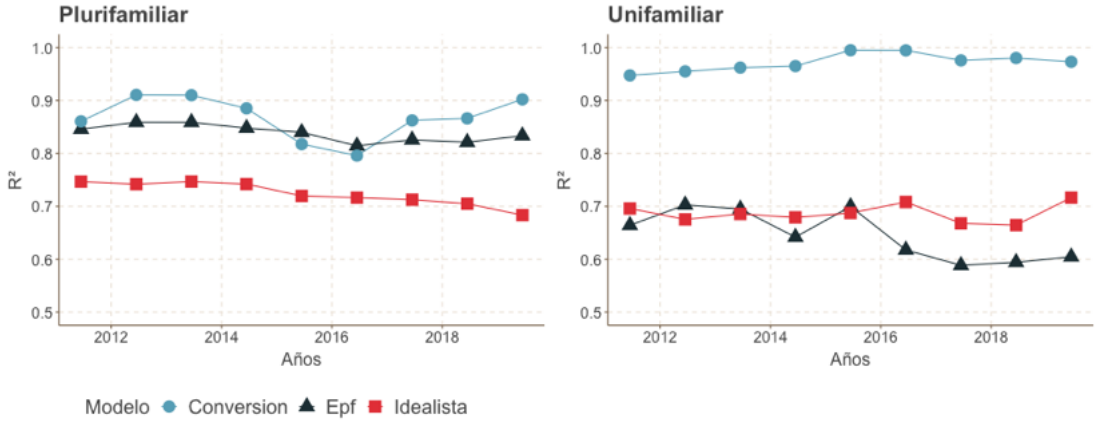
$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \qquad [3.11]$$

donde n son grados de libertad o número de observaciones, y k el número de regresores independientes.

[30] La calidad del ajuste representa la capacidad del modelo a predecir cualquier valor de la distribución de valores de la variable objetivo.

En los modelos lineales, el R^2 se estima de forma directa, mientras que en los de tipo *Random Forests*, se calcula sobre la muestra *out of bag*[31]. Esta última, es un conjunto de observaciones no utilizadas en la construcción de cada árbol[32] .

Figura 3.14. Coeficiente de determinación R²



Fuente: elaboración propia.

La Figura 3.14 muestra los niveles de ajuste de los modelos en términos de R^2, observándose que los niveles de ajuste son muy altos (en general mayores a 0,7), en particular para los modelos de correspondencia en ambos tipos (con una media en torno a 0,9). También se aprecia que en los modelos de EPF e Idealista los valores más bajos se corresponden a las viviendas unifamiliares, en comparación con las plurifamiliares. El menor ajuste de las primeras se puede deber a la ausencia de variables importantes en el modelo, como por ejemplo, el tamaño de la parcela o la limitación máxima de la superficie a 300 m².

En todo caso, un valor del coeficiente de determinación superior al 0,9 es consistente con los resultados de otros autores que utilizan métodos no paramétricos, como por ejemplo, el caso de Ho (2021) para Hong Kong o el Rico y Taltavull (2021) sobre precios de tasaciones en Alicante. Es importante destacar el grado de ajuste de nuestro caso puede considerarse muy bueno, dado que el número de atributos utilizado para la construcción del modelo de mercado es mucho más limitado que el de las publicaciones citadas.

La Tabla 3.9 muestra con mayor detalle los valores de R^2 obtenidos al ajustar los modelos.

[31] Para más detalle sobre el proceso de muestreo, véase el Anexo 3b del presente capítulo.
[32] En el proceso de *bagging* se divide el conjunto de datos en dos partes, *in bag* que se refiere a las instancias usadas para entrenar el modelo, y *out of bag*, en adelante OOB, que se usa para medir el ajuste y error del modelo.

Tabla 3.9. Resumen de ajuste de los modelos de mercado

	Plurifamiliar			Unifamiliar		
Año	**EPF**	**Idealista**	**Corresp.**	**EPF**	**Idealista**	**Corresp.**
2011	0,85	0,75	0,86	0,66	0,70	0,95
2012	0,86	0,74	0,91	0,70	0,68	0,95
2013	0,86	0,75	0,91	0,70	0,69	0,96
2014	0,85	0,74	0,89	0,64	0,68	0,97
2015	0,84	0,72	0,82	0,70	0,69	1,00
2016	0,81	0,72	0,80	0,62	0,71	0,99
2017	0,83	0,71	0,86	0,59	0,67	0,98
2018	0,82	0,70	0,87	0,59	0,66	0,98
2019	0,83	0,68	0,90	0,60	0,72	0,97

Fuente: elaboración propia

Para facilitar la lectura de los resultados de los modelos GAM de correspondencias, se han construido las Tablas 3.11 y 3.10, en las que se representan el signo[33] del coeficiente con su significatividad[34] . En ellas se observa que aquellos coeficientes que son más significativos suelen mantener consistencia de signo a lo largo del tiempo . Además, las dos tipologías muestran comportamientos muy diferentes, las viviendas plurifamiliares cuentan con un mayor grado de significatividad en sus coeficientes y consistencia temporal en los signos de los coeficientes .

Para el caso de las viviendas unifamiliares (Tabla 3.10) la consistencia en términos de signo y significatividad es menor, debido a tener una muestra más pequeña e inestable (véase epígrafe 2.4.3). En términos de mayor estabilidad temporal se pueden destacar los coeficientes de tipo de zona y los de número de habitaciones. Por otra parte, se observan diferencias a lo largo del tiempo, por ejemplo, los años 2015 y 2016 muestran niveles de significatividad mayores que en 2013, 2014 y 2017, en los que los coeficientes no son significativos. Es particularmente interesante destacar que aún así estos tres periodos tienen un R^2 era mayor a 0,9.

Adicionalmente, en el Anexo 3e se adjuntan ejemplos de los coeficientes de los modelos GAM de la EPF y oferta para viviendas unifamiliares. Se observa que el modelo sobre la EPF es mucho más débil en términos de significatividad, siendo las covariables más representativas: el tipo de edificio, tamaño del municipio, número de habitaciones, comunidad autónoma, si está o no en la capital de provincia, y un muy alto nivel de gasto familiar.

[33] "+" expresa que el signo del coeficiente es positivo y "-" que es negativo
[34] Se representan en función de p-valor: *** < 0.001, ** < 0.01, * < 0.05 y "." < 0.1

Tabla 3.10. Signo y significancia de coeficientes modelo GAM correspondencia para viviendas unifamiliares

	Signo									Significatividad								
Coeficiente	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
INTERCEPT	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	-	+	-	+	+	-	+	+	+				***	***		**	***	*
TAMAMUMunicipio con 20.000 o más y menos de 50.000	+	-	-	+	+	+	+	+	+	**			***	**		**	***	*
TAMAMUMunicipio con 10.000 o más y menos de 20.000	-	-	-	+	-	-	-	-	-	***	***	**	***	**	***	**		***
TAMAMUMunicipio con menos de 10.000 habitantes	-	-	-	+	-	+	-	-	-		***	.	***	***		***	***	**
TIPOEDIFVivienda unifamiliar adosada o pareada	-	+	-	+	+	-	+	+	-			**	**		***	**	***	.
ZONARESUrbana alta	+	+	+	-	+	+	+	-	+			**	*		***		***	
ZONARESUrbana media	-	+	+	-	-	+	-	-	-	*			**	***	***		***	.
ZONARESUrbana inferior	-	+	+	-	+	+	-	-	-	*		*	.	***	***		***	**
ANNOCONHace 25 ó más años	-	-	-	-	-	-	-	+	+		*	***	**	***	***	***	***	***
DENSIZona intermedia	-	+	-	-	-	+	-	+	+	***			***		***		***	***
DENSIZona diseminada	-	-	-	-	-	-	-	-	+	***	***	.	***	***	*	***	***	
INTERINPSPDe 500 a menos de 1000 €	+	+	-	+	+	-	+	+	-						***	***	***	
INTERINPSPDe 1000 a menos de 1500 €	+	+	-	+	+	-	+	+	+		*		**	***	***	***	***	
INTERINPSPDe 1500 a menos de 2000 €	+	+	-	+	+	-	+	+	+		.		***	***	***	***	***	
INTERINPSPDe 2000 a menos de 2500 €	-	+	-	+	+	-	+	+	-		*		**	***	***	***	***	
INTERINPSPDe 2500 a menos de 3000 €	-	+	-	+	+	-	+	+	+			**	**	***	***	**	**	*
INTERINPSP3000 o más €	-	+	-	+	+	-	+	+	+			**	*	**	***	**		**
NHABIT3 habitaciones	-	-	-	-	-	-	+	+	+	*	***	***		***	***	**	***	***
NHABIT4 habitaciones	-	-	-	+	-	-	+	+	+		***	***	*	***	***		***	**
NHABIT5 o más habitaciones	-	-	-	+	+	-	+	+	+	***	***	***			***			
CAPROVNo	+	-	+	-	-	-	-	-	-		**		***	***	***	***	***	***
factorGASTOT_1	-	-	-	+	+	-	+	+	-			*	***	***	***	***		
factorGASTOT_2	+	-	+	+	+	-	+	+	+	**	.		***	***	***	***	**	
factorGASTOT_3	+	+	+	+	+	-	+	+	+	**			***	***	***	***	*	
factorGASTOT_4		+	+	+	+						***	.	***	***				

Fuente: elaboración propia

Tabla 3.11. Signo y significancia de coeficientes modelo GAM correspondencia para viviendas plurifamiliares

	Signo									Significatividad								
Coeficiente	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
INTERCEPT	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 10.000 o más y menos de 20.000	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***		***	***
TAMAMUMunicipio con menos de 10.000 habitantes	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TIPOEDIFCon 10 ó más viviendas	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TIPOCASACasa económica o alojamiento	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
ZONARESUrbana alta	+	-	+	+	-	-	-	-	-	***	***	***		***	***	***	***	***
ZONARESUrbana media	-	-	-	-	-	-	-	-	-		***	*	***	***	***	***	***	***
ZONARESUrbana inferior	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
ANNOCONHace 25 ó más años	+	+	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
DENSIZona intermedia	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
DENSIZona diseminada	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
INTERINPSPDe 500 a menos de 1000 €	-	+	-	+	-	-	-	+	-		***	***	***	***	***	***	***	***
INTERINPSPDe 1000 a menos de 1500 €	+	+	+	+	-	-	-	+	-	***	***	***	***	***	***	***	***	
INTERINPSPDe 1500 a menos de 2000 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	*	***	***
INTERINPSPDe 2000 a menos de 2500 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
INTERINPSPDe 2500 a menos de 3000 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
INTERINPSP3000 o más €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
NHABIT3 habitaciones	-	-	-	-	-	-	-	+	+	***	***	***	***	***	***	***	***	***
NHABIT4 habitaciones	-	+	+	+	-	-	-	-	-	***	***	***	***	***	***	***	***	**
NHABIT5 o más habitaciones	-	+	-	+	-	+	-	-	-	***	***		***	**		***	***	
CAPROVNo	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
factorGASTOT_1	-	+	+	-	+	-	+	-	+	***	***	***	***	***	*		***	***
factorGASTOT_2	+	+	+	-	+	+	+	-	+	***	***	***	***	***	***		***	***
factorGASTOT_3	+	+	+	+	+	+	+	-	+	***	***	***		***	***		***	***
factorGASTOT_4		+	+	+	+	+	+	-			***	***	***	***	***	*	***	

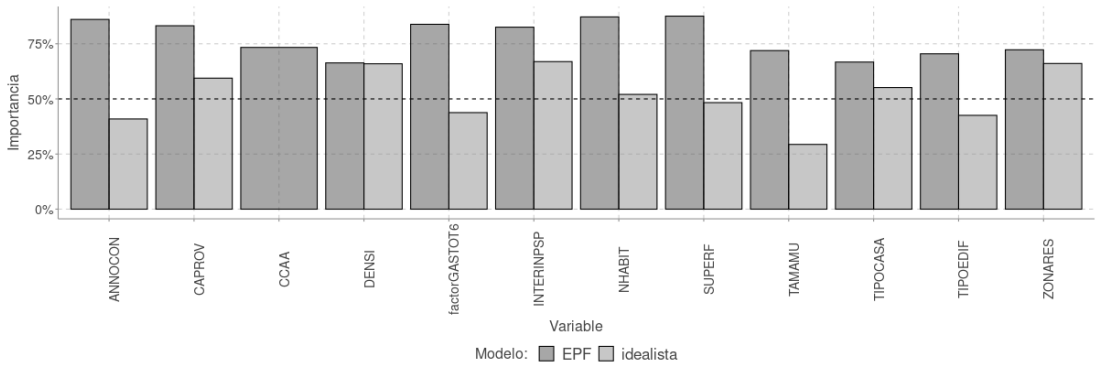
Fuente: elaboración propia

Dado que para los modelos construidos con *Random Forests* no es posible estimar una significatividad de los coeficientes, se puede analizar la contribución de las variables en función de su capacidad de eliminación de la entropía o "impureza"[35]. Esta medida se puede entender como la importancia de un predictor para reducir la variabilidad de los residuos del modelo (que por otra parte es el principio sobre el que se basa el R^2).

Tanto el peso como la significatividad de los coeficientes permiten estudiar el efecto de las variables auxiliares en el proceso de correspondencia en los modelos de tipo no paramétrico. Es conveniente analizar ambas medidas, ya que generalmente su eficacia depende del caso (Zhang y Nguyen, 2020).

Por otra parte, en Figura 3.15 se muestra la importancia de las 12 variables más importantes. Para cada una de las variables del modelo de la EPF se indica su nivel de importancia normalizada con respecto a su aporte en el modelo[36]. Se observa que todos los casos los valores son muy altos, lo que indica que son significativas e intervienen casi en la misma medida en la construcción de los árboles de decisión.

Figura 3.15. Importancia de las variables para los modelos EPF e idealista para viviendas plurifamiliares



Fuente: elaboración propia.

En el modelo de la EPF destacan como variables con más peso: el año de construcción, la provincia, el número de habitantes del municipio, la provincia y los factores de gastos e ingresos. Para el modelo de oferta de Idealista, las variables más importantes son la provincia, densidad de población, nivel de ingresos y tipo de zona residencial, confirmando la importancia de la zona en el modelo hedónico.

A la vista de los resultados de ajuste y de importancia de variables, se puede concluir que la selección de variables auxiliares para la calibración cumple los

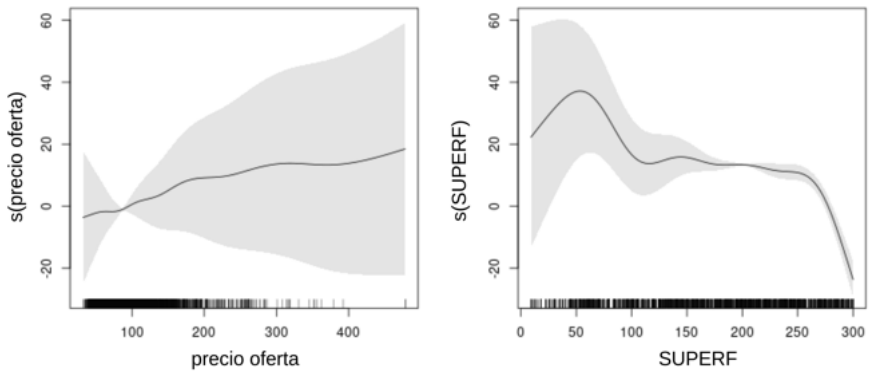
[35] En la configuración utilizada se ha utilizado *impurity* como medida de importancia de variables.

[36] La importancia se calcula se calcula en términos de "impureza" y representa la varianza que reducen los cortes en los que interviene una variable, este valor se normaliza diviendolo por la máxima medida de impureza entre todas las variables.

tres criterios de Särndal y Lundström (2008): 1) explicar la variable respuesta; 2) que las covariables sean todas significativas; y 3), servir para desarrollar la estratificación del índice de precios.

En importante resaltar que los intervalos de confianza de las funciones de suavizado de los modelos GAM, en las viviendas unifamiliares, son muy amplios (representados en la Figura 3.16 el color gris alrededor de la línea de regresión). Lo cual puede relacionarse un menor nivel de grado de R^2. En el caso particular de la superficie útil (*SUPERF*), el intervalo es muy amplio en los valores más bajos, lo que indica que este factor es poco representativo en los inmuebles más pequeños.

Figura 3.16. Relación de valor de la función de suavizado (s) con las variables precio de oferta y superficie útil, para el modelo de correspondencia en vivienda unifamiliar

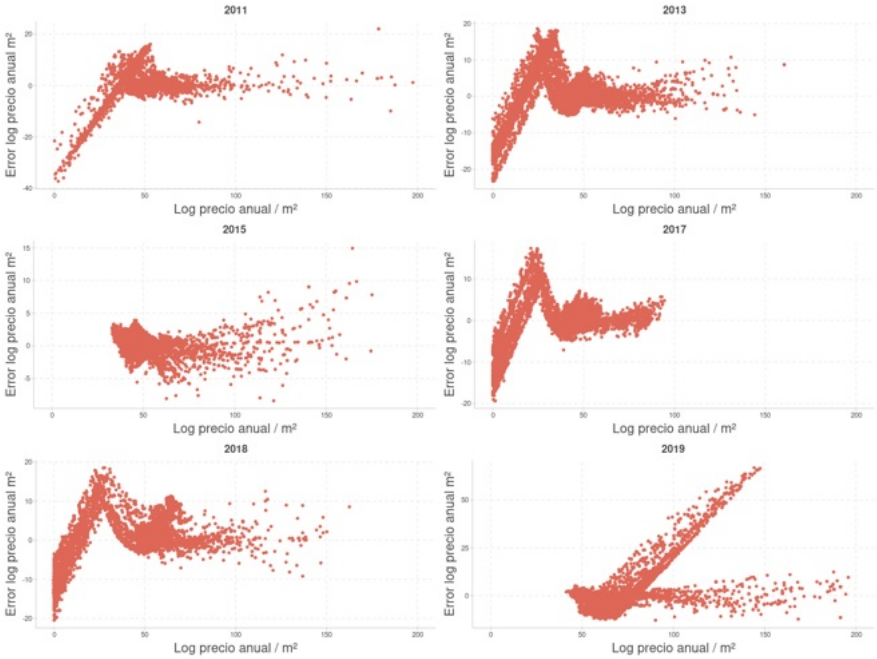


Fuente: elaboración propia.

Por otra parte, el precio de oferta (*preciom2_anualpred*) muestra una amplia variabilidad en todo el rango de valores, a excepción de los valores cercanos a 100 €/m²/año. En los valores superiores, el motivo podría ser la existencia de variables omitidas importantes como son el tamaño de la parcela, el estado de conservación de la vivienda, o la limitación de superficie útil a un máximo de 300 m². La incertidumbre en los valores bajos (menos de 100 m²) es menos importante porque para este tipo de propiedad son áreas infrecuentes.

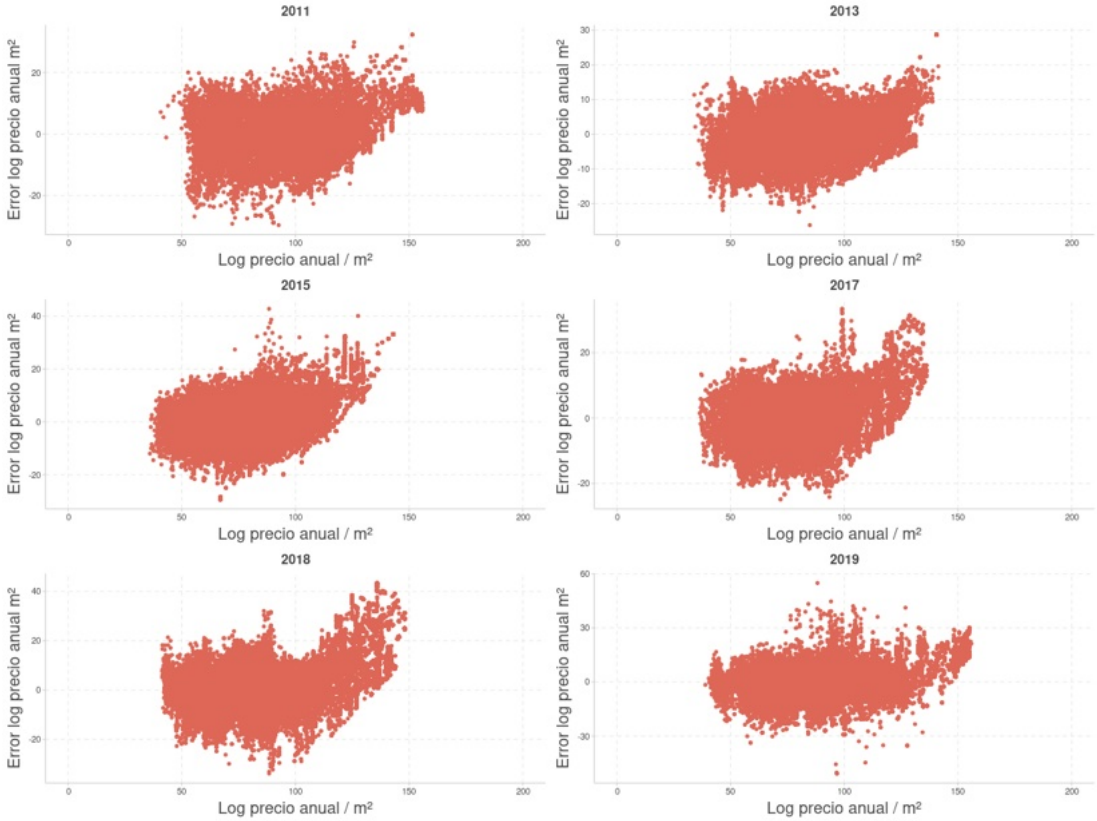
Para confirmar la hipótesis anterior, se representan los errores del modelo de viviendas unifamiliares en la Figura 3.17. Se aprecia como los errores muestran un patrón estable en el tiempo, excepto para los casos de 2015 y 2019 cuando ofrecen un patrón menos definido. En los primeros casos, los errores son mínimos para los precios más bajos, y que ascienden hasta un punto donde vuelven a descender y estabilizarse. Este comportamiento inestable de los residuos es habitual en muestras pequeñas y muy heterogéneas (Goh *et al.*, 2012), como la del segmento de estudio.

Figura 3.17. Residuos modelo de correspondencia en escala logarítmica, vivienda unifamiliar



Fuente: elaboración propia.

Figura 3.18. Residuos modelo de correspondencia en escala logarítmica, vivienda plurifamiliar

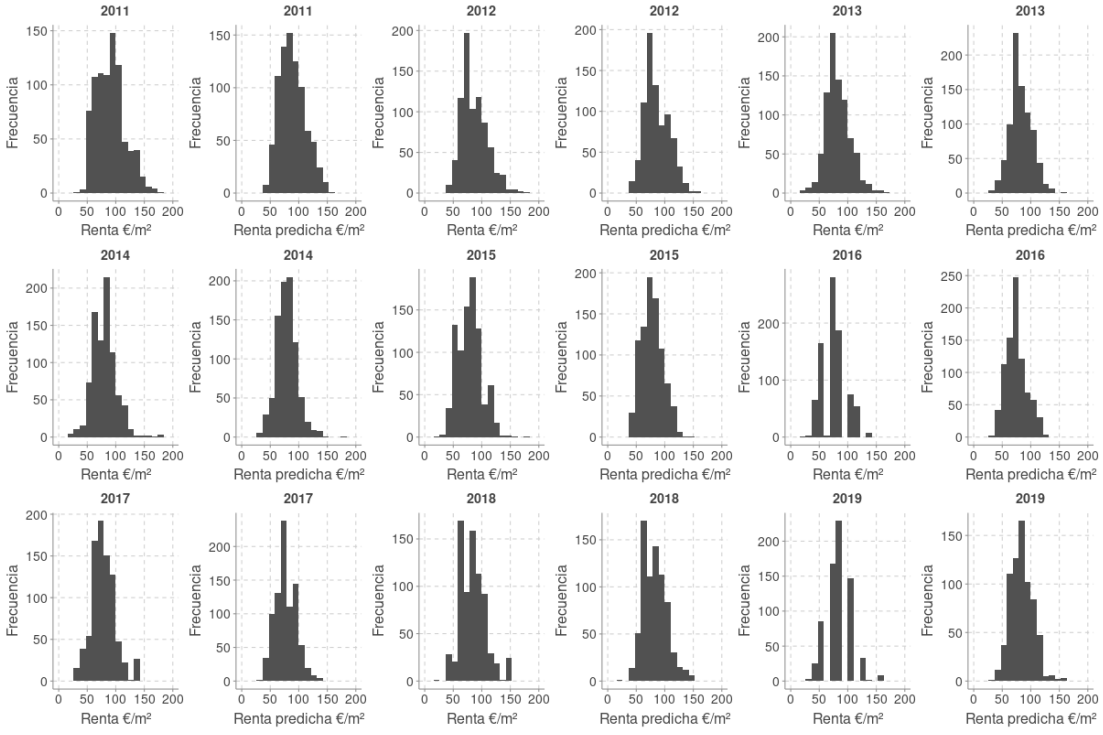


Fuente: elaboración propia.

Esto contrasta con los errores en el modelo para viviendas plurifamiliares de la Figura 3.18, donde los residuos, en escala logarítmica, están más cercanos a un patrón aleatorio. Además, la distribución de los valores son consistentes a lo largo del tiempo, y se aprecia que los errores son mayores en los precios más altos. Esto se debe a que los segmentos más caros contienen una mayor proporción de inmuebles singulares, cuyos precios no siguen el patrón general, y por tanto, el grado de imprecisión de los modelos en este tramo de precios es mayor.

Por último, debe comprobarse que la distribución de la variable de interés se mantiene en los ficheros donantes y receptores (Rässler, 2012). Para ello, primeramente, se evalúa si el modelo de imputación la EPF lo hace. En la Figura 3.19 se muestran las distribuciones del precio del alquiler original y estimada. Se observa que el modelo mantiene la forma de las distribuciones originales, con dos excepciones: en 2017 la predicción tiende a concentrar los valores en torno a la mediana; y en los años 2016 y 2019 los ficheros originales no ofrecen una forma continua cuando el modelo si lo hace.

Figura 3.19. Distribución original y predicciones para el modelo de imputación de valores de la EPF



Fuente: elaboración propia.

A tenor de lo anterior, la condición de preservación de las distribuciones marginales se considera válida al existir convergencia en los procesos de calibración.

3.4.3 Distribución conjunta

Para comprobar que distribución conjunta del proceso de correspondencia se mantiene, se debe asegurar que la distribución del modelo mantiene las propiedades del fichero de la EPF. En lo que se refiere a la distribución de frecuencias, existen diversas formas de medir el nivel de divergencia entre poblaciones, una de las más utilizadas (Leucescu y Agafitei, 2013) es la distancia Hellinger $H_d(P,Q)$, que puede aplicarse tanto para poblaciones continuas como discretas. Existen otras alternativas como las pruebas Chi-cuadrado, Kolmogorov Smirnov, Rao-Scott, Wald-Wolfowitz, que se estudian en detalle en Corder y Foreman (2014).

En este caso, la información de la EPF se conoce de forma discreta, agrupada por estratos (*YEAR*, *TAMAMU, TIPOEDIF*, *TIPOCASA*, *ZONARES*, *DENSI*, *ANNOCON*, *INTERINPSP*, *factorGASTOT6*, *NHABIT*, *CAPROV*), por tanto, se utiliza una variante de la distancia Hellinger aplicada a poblaciones discretas, calculada según la siguiente expresión analítica:

$$H_d(P,Q) = \frac{1}{n_d} \cdot \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2} \qquad [3.12]$$

donde n_d es el número de dimensiones, p_i es la probabilidad de inclusión del estrato i para la tabla de contingencia de los pesos del modelo, que en este caso serán los pesos calculados por la calibración, y q_i es la correspondiente probabilidad para el mismo estrato para la EPF. Los estratos i se engloban en un conjunto K total de estratos, cuyas probabilidades de inclusión p_i y q_i se definen como:

$$p_i = \frac{n_i}{N}, q_i = \frac{n'_i}{N'} \qquad [3.13]$$

donde N es el total de la tabla de contingencia, n_i la frecuencia tiene el estrato i, es decir el peso poblacional de este estrato en el conjunto P. Los n'_i y N' serían los respectivos valores para la EPF.

En este caso, se obtiene una distancia de Hellinger de 0,042 sobre un total de 2.723 estratos y 11 dimensiones, con una media de 248 estratos por dimensión. Según Leucescu y Agafitei (2013) se considera que esta distancia representa que dos distribuciones similares cuando su valor es menor de 0,05[37]. Se puede concluir, por tanto, que se mantiene la distribución conjunta de pesos poblacionales, al ser la distancia del mismo orden de magnitud que las obtenidas por los mismos

[37] Es cierto que esta medida debe usarse con precaución porque no tiene en consideración la variabilidad debida al diseño del muestro o cuando existen un gran número de categorías.

autores (Leucescu y Agafitei, 2013) en las armonizaciones de las encuestas de condiciones de vida (EU-SILC) y de población activa europea (EU-LFS). En cuyos casos obtuvieron un valor cercano al 0,04.

Para la distribución conjunta de los precios, se ha comprobado que los parámetros principales del fichero de la EPF se mantienen en el fichero definitivo, tanto en órdenes de magnitud como en tenencia. La Tabla 3.12 muestra estos parámetros en ambos ficheros (media, cuantiles y desviación estándar ponderados[38]). Se observa que la desviación típica en el fichero definitivo es ligeramente menor, a excepción del año 2014 cuando se produce un fuerte aumento de la desviación. Los valores absolutos de la media y cortes de cuantiles son superiores debido al efecto del suavizado exponencial aplicado a la EPF.

Tabla 3.12. Parámetros sobre precio €/m²/año población original y final

	Fichero EPF					Fichero final				
Año	**Media**	**Dev**	**Q1**	**Q2**	**Q3**	**Media**	**Dev**	**Q1**	**Q2**	**Q3**
2011	92	657	71	90	106	100	599	81	101	119
2012	88	569	72	83	103	96	609	76	99	115
2013	84	441	70	80	97	91	474	74	93	109
2014	81	507	68	80	91	90	803	71	87	106
2015	79	435	63	80	90	86	398	72	87	101
2016	76	437	58	75	88	83	375	71	82	98
2017	79	442	58	75	95	86	417	74	86	100
2018	83	559	62	84	95	90	562	72	89	105
2019	85	512	70	87	103	95	549	76	96	110

Fuente: elaboración propia

Aún cuando los el nivel de ajuste es muy alto, se puede comprobar gráficamente que las medidas promedio de la variable de interés varían sensiblemente en función de los elevadores muestrales utilizados, a pesar de referirse al mismo colectivo. Lo cual puede comprobarse en la Figura 3.20, dónde se representa la suma de los precios ponderados de los estratos individuales[39] usando dos conjuntos de pesos: los originales de la EPF y los estimados por el proceso de calibración ($g \cdot w$).

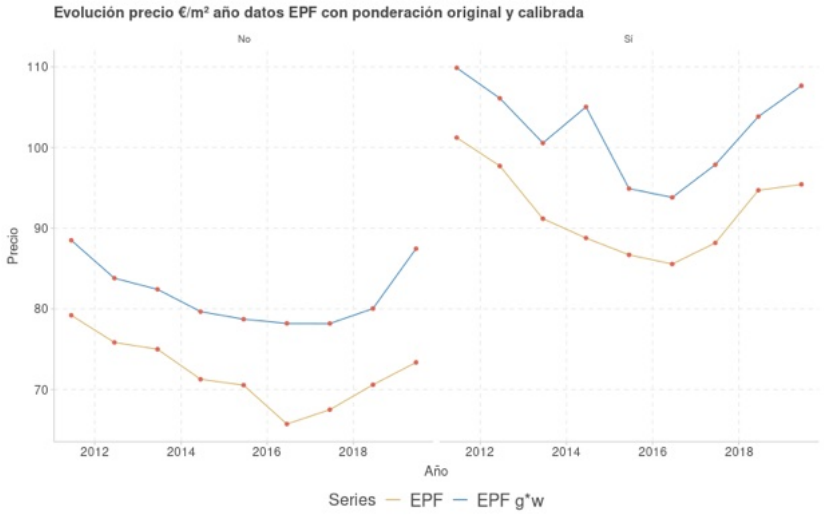
La Figura muestra que la serie de precios basada en la calibración tiene comportamientos anómalo, desde un punto de vista de lógica de mercado. Por

[38] En cada fichero se usan pesos diferentes, para la EPF los pesos del fichero original y en el fichero de oferta se usan los factores de elevación procedentes de la calibración.

[39] Se parte de las celdas de menor tamaño siguiendo la estratificación de la población mediante las covariables del modelo de correspondencia.

ejemplo, se observa un incremento importante de precios entre 2018 y 2019 para el resto de Comunidad de Madrid, y una subida puntual del precio en 2014 paras la ciudad de Madrid, ambos valores no se corresponden a ninguna causa de mercado justificable.

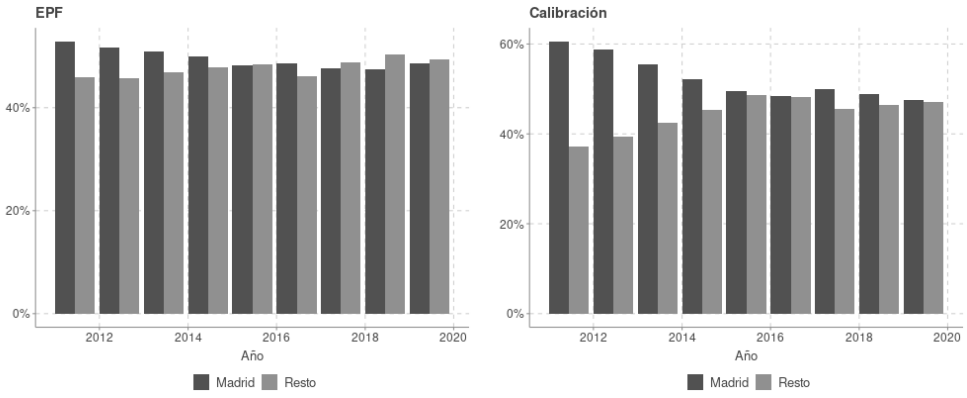
Figura 3.20. Series de precio promedio: pesos EPF y pesos calibrados



Fuente: elaboración propia.

Las diferencias en las tendencias de la figura anterior se deben a un efecto de composición con los nuevos pesos, que lógicamente difieren de los originales. La Figura 3.21[40] muestra las diferencia entre los pesos originales y los calibrados para los estratos definidos.

Figura 3.21. Pesos poblacionales EPF y calibración por Madrid o resto de zonas



Fuente: elaboración propia.

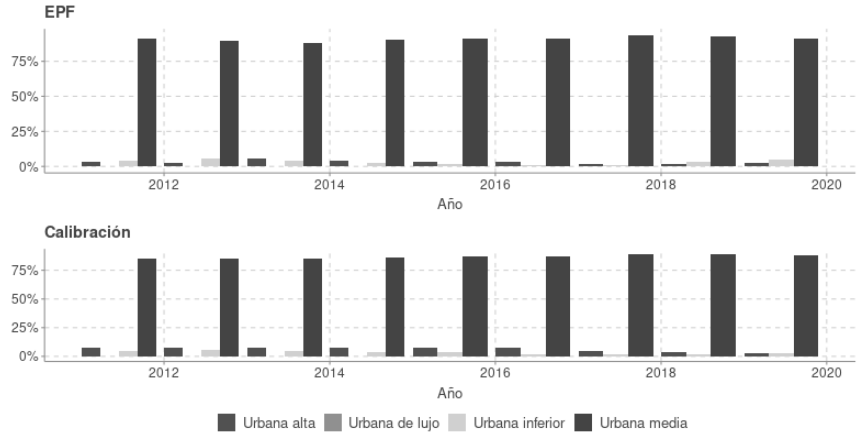
Los pesos de la EPF son relativamente estables a lo largo del tiempo. En cambio, los pesos calibrados parten de una situación mucho más desequilibrada en 2011

[40] La variable *CAPROV* indica si la observación se encuentra en la capital de provincia.

que en el 2019.

Para otros criterios de estratificación, como el tipo de zonas residencial, no se aprecian unas diferencias tan acusadas. La Figura 3.22 que muestra ambas distribuciones de pesos, indica que existe una mayor representación de las zonas minoritarias en la calibración, pero la desigualdad es mínima.

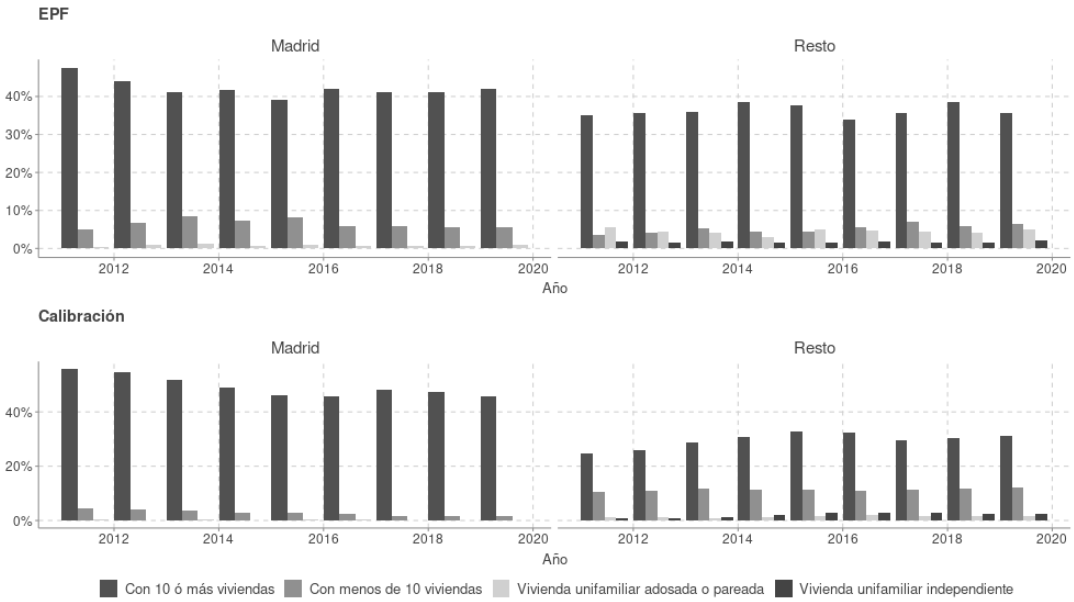
Figura 3.22. Pesos poblacionales EPF y calibración por tipo de zona residencial



Fuente: elaboración propia.

En el caso del desglose por edificio, hay una mayor distribución por tipo en la ciudad de Madrid, y una mayor diversidad de tipos en el resto de provincia en la población calibrada por el censo, como se observa en la Figura 3.23.

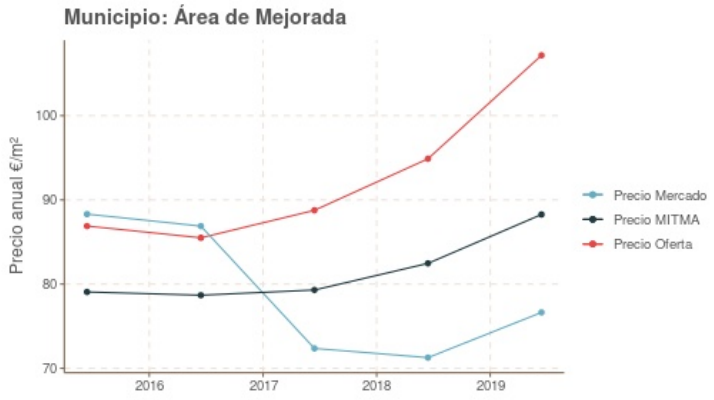
Figura 3.23. Pesos poblacionales EPF y calibración por tipo de edificio



Fuente: elaboración propia.

Las medidas anteriores no tienen en cuenta el desglose zonal, ya que el dato a este nivel no está disponible en el fichero “donante” de la EPF y no puede utilizar como variable común para hacer el enlace. Se puede anticipar que la falta de referencias geográficas genera sesgos en las estimaciones de precios de zona en el modelo, a la vista de los valores del ejemplo de la Figura 3.24, para las series de precios en el área de Mejorada del Campo (zona en la que se aprecia de forma muy acusada este efecto). En la cual se observa una caída del precio de mercado entre 2015 y 2018, que no guarda coherencia con el precio registrado oficialmente por MITMA en ese municipio.

Figura 3.24. Precios para vivienda plurifamiliar, área de Mejorada del Campo



Fuente: elaboración propia.

Se puede apreciar también que la misma inconsistencia con MITMA se produce con los precios de oferta. El motivo de la misma podría deberse a que el modelo de mercado no ha tenido en cuenta la zona geográfica específica, al contrario que en oferta o MITMA. Por tanto se puede asumir que se introduce un sesgo por una insuficiente especificación zonal de los modelos hedónicos de mercado (al no disponerse de esta información en la EPF).

Para comprobar lo anterior, la Tabla 3.13 muestra dos métricas de discrepancia entre las series de oferta y mercado con respecto a las de MITMA en todas las zonas, entre 2015 y 2019. La primera medida (divergencia) compara el signo de las variaciones de cada una de las series, y expresa el número de veces en los que el signo de la variación difiere (normalizado por el número de observaciones). La segunda métrica (diferencia) representa la desviación media en porcentaje, entre las variaciones de la serie y la de MITMA.

Tabla 3.13. Divergencia variación anual de precios respecto a MITMA

		% divergencia		% diferencia	
Tipo	**N**	**Oferta**	**Mercado**	**Oferta**	**Mercado**
Plurifamiliar	668	7,2%	32,0%	5,1%	27,9%
Unifamiliar	236	40,3%	45,3%	12,2%	41,6%

Fuente: elaboración propia

Los resultados muestran que se puede generalizar la anomalía observada en Mejorada del Campo para el resto de las zonas, donde existe una divergencia alta en las series de mercado generadas por el modelo (superiores al 30%). En cambio, la coincidencia para la oferta es muy alta en términos de signo, especialmente para las viviendas plurifamiliares, con una divergencia del 7,1% y una desviación en términos absolutos de un 5,1%, lo que hace suponer que estas series están altamente correlacionadas.

Las viviendas unifamiliares muestran un comportamiento más irregular, siendo las tasas de coincidencia en signo altas en todos los casos.

La causa principal del comportamiento anterior procede de la heterogeneidad espacial en las distribuciones de precios, que por otra parte, es un fenómeno conocido y ampliamente documentado. Entre otros autores, la cuestión ha sido analizada en profundidad por Hu (2022), Wu (2020), Helbich (2014), Páez (2008) y Kestens (2006). Para nuestro caso, se pueden identificar dos orígenes de la heterogeneidad:

- La calibración final con la EPF no tiene información zonal.
- Ninguno de los modelos de mercado se refieren a zonas concretas, sino que son estratos de tipo funcional o de características[41].

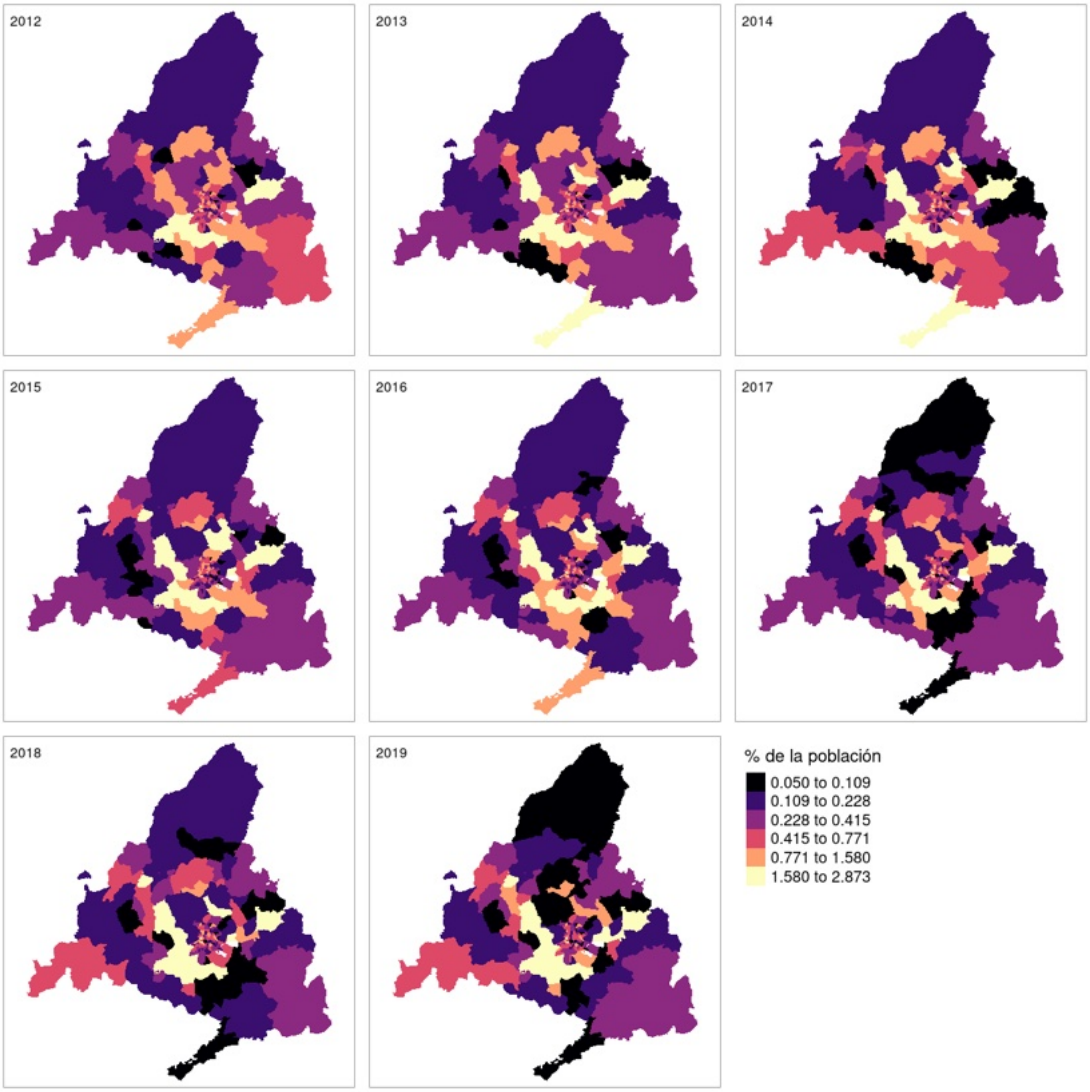
[41] Incluso las características de clasificación de zona: rural, urbana, densamente poblada o según ingresos no recogen las diferencias entre zonas

3.4.4 Distribución espacial de la población

Es requisito indispensable que la distribución de los pesos muestre estabilidad temporal, y se corresponda la de la poblacional real. De la misma manera, y aunque se parta de la limitación del desconocimiento de las distribuciones exactas del alquiler por zona geográfica, el proceso de calibración ideal debería ser capaz de replicar la distribución zonal de oferta en la medida de lo posible[42].

En el epígrafe anterior, los niveles de ajuste del modelo confirman que el modelo replica el desglose funcional de la muestra. Por tanto, si es necesario un ajuste, será para garantizar que se mantiene la coherencia zonal de la información, manteniendo el comportamiento funcional actual.

Figura 3.25. Distribución espacio-temporal de los pesos poblacionales en vivienda plurifamiliar, toda la Comunidad de Madrid



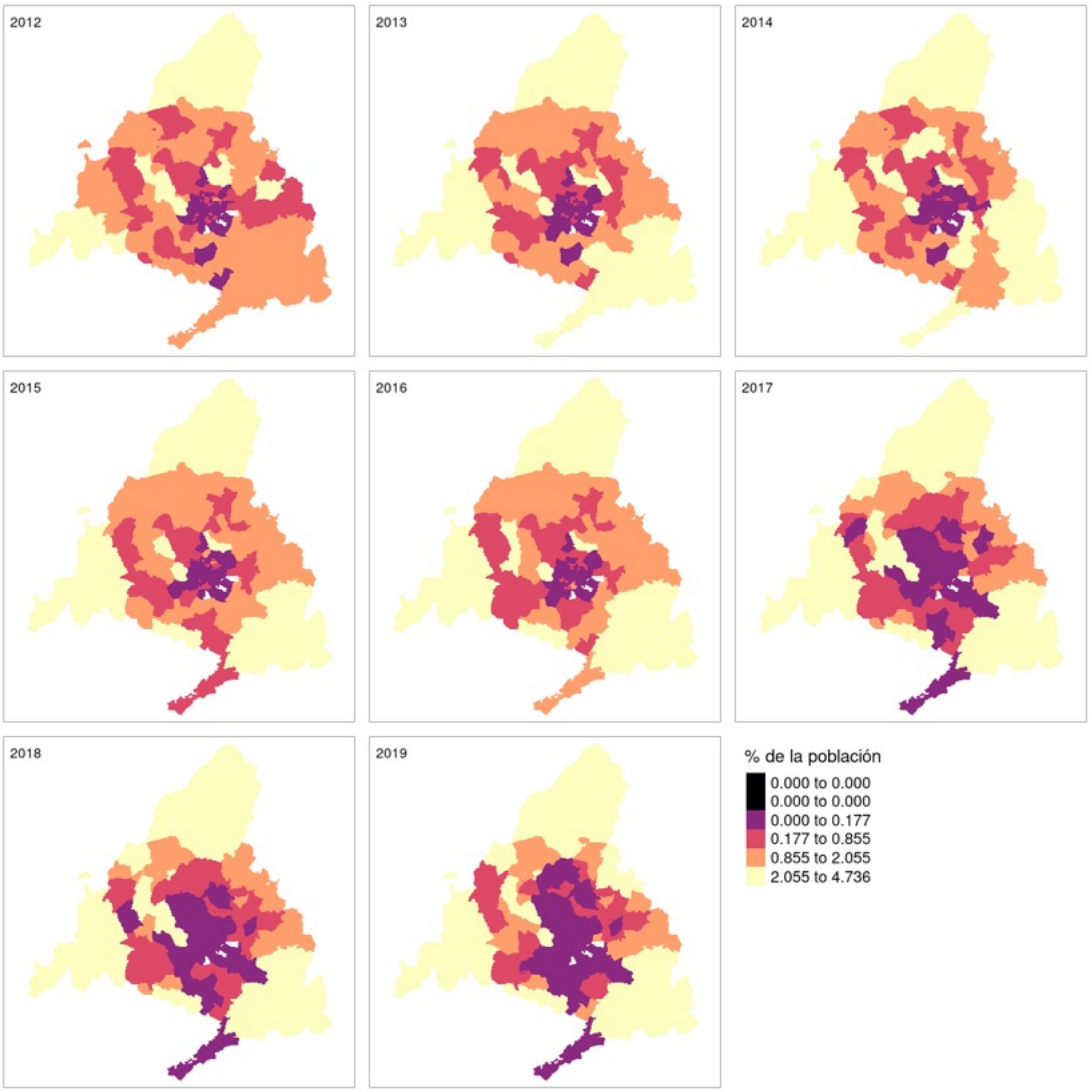
Fuente: elaboración propia.

[42] Entendiendo que en ausencia de información de las rentas reales, el dato de oferta ofrece una medida aproximada, aunque sea en términos de orden de magnitud y evolución temporal.

Para las viviendas plurifamiliares, se observa en la Figura 3.25, que la representación de las zonas rurales decrece en el tiempo, mientras que, la zonas metropolitanas sur y oeste ganan progresivamente importancia en la muestra. Existe una mayor variabilidad en el tiempo en las zonas centrales de la Comunidad, con respecto a las zonas exteriores.

Las viviendas unifamiliares, en cambio, se concentran en la corona justamente exterior al centro, ampliándose el radio interior de forma progresiva, como vemos en la Figura 3.26. Esto indica que la población de alquiler de las zonas metropolitanas exteriores ganan progresivamente más importancia.

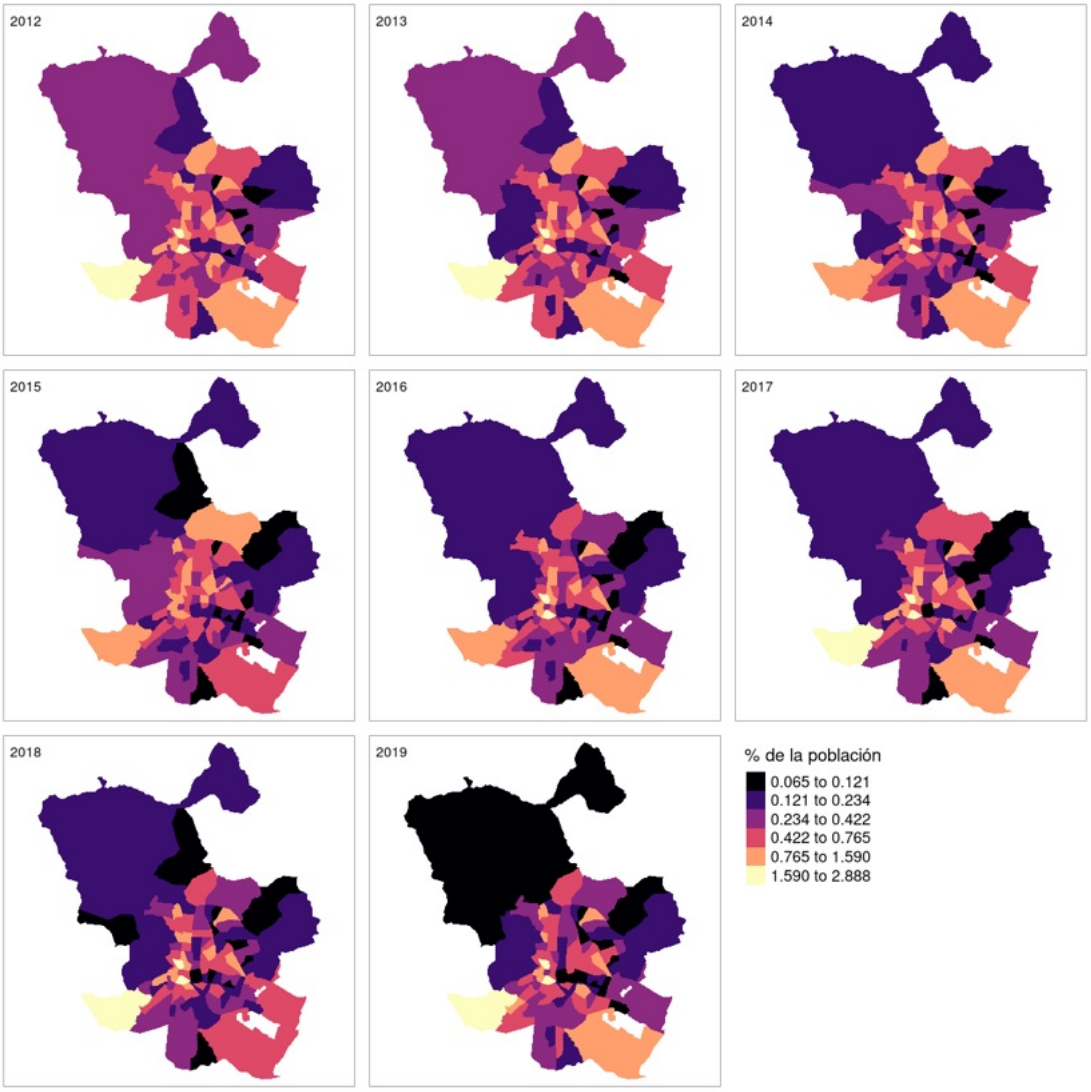
Figura 3.26. Distribución espacio-temporal de los pesos poblacionales en vivienda unifamiliar, toda la Comunidad de Madrid



Fuente: elaboración propia.

Para el caso de la ciudad de Madrid, se observa que el peso de las distintas zonas varía en el tiempo, como se aprecia en la Figura 3.27, aunque son la zona central y la sur las que mantienen más registros en términos relativos.

Figura 3.27. Distribución espacio-temporal de los pesos poblacionales para todos los tipos, ciudad de Madrid



Fuente: elaboración propia.

La Tabla 3.14 muestra las zonas que más han variado en términos relativos. Destaca la zona metropolitana del noroeste como aquella con variaciones más altas a partir del año 2017. Por ejemplo, Majadahonda pasa de un 1,1% del total de la población al 4,8%. En general, estos municipios del noroeste (Majadahonda, Las Rozas, Collado Villalba y Pozuelo de Alarcón) representaban un 4,2% en 2011, mientras que en 2019 cuentan con un 14,4% de las viviendas en alquiler. Estas áreas conforman el eje de mayores ingresos de la Comunidad, y también, cuentan con un nivel de precios por superficie mayor a la media.

Tabla 3.14. Zonas con mayor variación en porcentaje

Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Majadahonda	1,1%	1,4%	1,6%	2,0%	2,6%	2,7%	4,3%	4,7%	4,8%
Rozas de Madrid, Las	1,1%	1,4%	1,7%	2,1%	2,7%	2,7%	4,2%	4,5%	4,5%
Coslada	0,9%	1,1%	1,3%	1,5%	2,3%	2,4%	3,8%	3,9%	4,1%
Pozuelo de Alarcón	1,0%	1,3%	1,4%	1,8%	2,5%	2,3%	3,5%	3,7%	3,7%
Collado Villalba	1,0%	1,2%	1,3%	1,5%	2,3%	2,3%	3,6%	3,6%	3,4%
Alcalá de Henares	2,9%	3,1%	3,1%	3,3%	2,8%	2,6%	2,2%	2,2%	2,2%
Móstoles	2,0%	1,9%	2,0%	1,9%	1,9%	1,9%	1,9%	1,9%	2,4%
Getafe	2,0%	2,0%	1,9%	1,7%	1,8%	1,8%	1,5%	1,8%	1,6%
Embajadores	2,1%	1,9%	2,0%	1,8%	1,6%	2,2%	2,1%	2,4%	2,4%
Universidad	1,7%	1,8%	1,7%	1,8%	1,3%	1,7%	1,7%	2,0%	1,8%

Fuente: elaboración propia

Las diferencias por tipología y capital de provincia (Tabla 3.15) muestran mayor estabilidad en Madrid (mediana del 0,11% contra un 0,19%). También existen menores diferencias en la vivienda plurifamiliar que en la unifamiliar (del 0,82% y un 0,15% respectivamente, para el resto de la CAM). Más concretamente, los cambios más importantes y fuente de variabilidad en los precios, se producen en viviendas unifamiliares fuera de la capital.

Tabla 3.15. Zonas con mayor variación anual media (periodo 2011 a 2019)

Capital	Tipo	Mín.	p05	p25	p50	p75	p95	Máx.
No	Todos	0,00%	0,02%	0,08%	0,19%	0,60%	2,61%	3,63%
	Plurifamiliar	0,01%	0,01%	0,06%	0,15%	0,57%	2,59%	3,85%
	Unifamiliar	0,01%	0,06%	0,32%	0,82%	1,34%	4,05%	10,74%
Sí	Todos	0,00%	0,01%	0,06%	0,11%	0,20%	0,31%	0,58%
	Plurifamiliar	0,00%	0,01%	0,06%	0,11%	0,20%	0,31%	0,56%
	Unifamiliar	0,00%	0,00%	0,00%	0,02%	0,13%	0,66%	1,31%

Fuente: elaboración propia

A lo largo de este capítulo se ha desarrollado un modelo de mercado que relaciona los precios de oferta con los de mercado, que sin embargo cuenta con limitaciones por la débil especificación zonal de la EPF. Esta cuestión se solventará con el modelo hedónico del capítulo 6, incorporando información zonal a través de medidas de las accesibilidad presentadas en el próximo capítulo.

Anexo 3a. Métodos de calibración

Aunque originalmente la calibración se orientó a la reducción de la varianza, se utiliza a menudo para corregir los sesgos muestrales. En contraposición con el enfoque clásico, en el que se construyen grupos de respuesta homogéneos[43], en la calibración, las variables de respuesta deben conocerse únicamente para las unidades que responden, junto con los totales poblacionales, lo que resulta bastante restrictivo. Para resolverlo, Deville desarrolló la teoría de calibración "supergeneralizada" (Deville, 2000), donde no es necesario conocer los totales poblacionales de las variables en las subpoblaciones sin respuesta, y se trabaja solo con las unidades cuya respuesta es conocida.

El método supone utilizar información auxiliar relacionada con la variable de estudio para ajustar los pesos de la muestra. Su base consiste en que, dada la característica Y de estudio, existe una serie de variables auxiliares X fuertemente relacionadas con Y, cuyos datos son conocidos o son fácilmente accesibles.

La calibración permite mejorar los estimadores de los parámetros poblacionales basados en técnicas tradicionales calculadas como métodos lineales, cuadráticos o por diseños muestrales complejos. Todo ello redunda en métodos con menor error, y permite trabajar con poblaciones con un menor tamaño. Deville y Särdnal (1992) demuestran, además, que cuanto más fuerte es la relación de las variables auxiliares con las variables de estudio, más precisa es la calibración.

La idea de la calibración parte de que dada una función de distancia d, que se establece entre los pesos iniciales y finales, se puede encontrar una razón a aplicar a los pesos originales π_k, denominada g_k, que cumple que la suma de sus distancias es mínima. La condición a minimizar por tanto sería:

$$\sum_{R_k=1}^{N} d(g_k, 1) \qquad [3.14]$$

donde las variables y_i son las variables de interés conocidas, tanto de forma agregada como a nivel individual. La condición anterior está sujeta a las restricciones de calibración:

$$\sum_{k=1}^{N} y_i = \sum_{R_k=1}^{N} \frac{g_i}{\pi_k} \cdot y_i \qquad [3.15]$$

De una forma más formal, dada una población U compuesta de N elementos

[43] La muestra se divide en celdas donde se supone que la distribución de la variable de la respuesta es uniforme. Por tanto, la tasa de respuesta es una estimación máxima en verosimilitud para esa distribución.

(k) distintos, identificados a través de sus etiquetas $i = 1, ..., N$, se parte de una serie de características de interés y_i asociadas con el elemento i que se conoce exactamente, y sin error, observando el elemento i.

Una muestra s es un subconjunto de U cuyos valores asociados de Y, identificados como $\{(k, y_k)\}$, que se seleccionan según con un diseño muestral específico que asigna una probabilidad conocida $p(s) > 0$ para todo $s \in S$. Siendo S el conjunto de las todas las posibles muestras s, y que cumple ques $\sum_{S \in S} p(s) = 1$. El total poblacional T de la variable Y se calcularía como:

$$T_Y = \sum_{k \in U} y_k \qquad [3.16]$$

La muestra cuenta con un vector de variables auxiliares $X = (X_1, ...X_J)$ que es perfectamente conocido, para todos los elementos de la población U. De tal forma que consideramos el estimador de Horvitz-Thompson[44] $\hat{T}_Y$ asociado como:

$$\hat{T}_{Y_\pi} = \sum_{k \in S} d_k \cdot y_k \qquad [3.17]$$

que pretende es modificar los pesos originales d_k, calculados como:

$$d_k = \frac{1}{\pi_k} \qquad [3.18]$$

por otros pesos ω_k, de forma que, el estimador basado en dichos pesos proporcione estimaciones perfectas para X, es decir, que cumpla:

$$\sum_s \omega_k X_k = T_X = (T_{X_1}, .., T_{X_J}) \qquad [3.19]$$

y estén tan próximos como sea posible, según una medida de distancia dada, a los pesos originales d_k. El método más común para definir esta distancia es la suma ponderada de los cuadrados de las distancias:

$$\sum_{k \in S} \frac{(\omega_k - d_k)^2}{q_k \cdot d_k} \qquad [3.20]$$

donde q_k son constantes positivas, que se resuelven como un problema de minimización de la expresión anterior, con la siguiente restricción:

[44] Un estimador de Horvitz-Thompson es un método para estimar el total y la media de una pseudopoblación en una muestra estratificada (Fuller, 2011; Särndal *et al.*, 2003).

$$\sum_{s} \omega_k \cdot X_k = T_X \quad [3.21]$$

usando el método de los multiplicadores de Lagrange, se obtienen los siguientes pesos, ya calibrados como:

$$\omega_k = d_k + d_k \cdot q_k \cdot \lambda X'_k \quad [3.22]$$

suponiendo que la inversa de $T_S = \sum_{k \in S} d_k \cdot q_k \cdot X_k \cdot X'_k$ existe, el estimador calibrado vendría dado como el estimador general de regresión, véase (Cassel *et al.*, 1976), definido según la expresión analítica:

$$\hat{T}_{Y_{reg}} = \sum_{k \in S} \omega_k y_k = \hat{T}_{Y_\pi} + (\hat{T}_X - \hat{T}_{X_\pi}) \cdot \hat{B}_S \quad [3.23]$$

La forma que el total estimado, $\hat{T}_{Y_{reg}}$, dependerá del diseño muestral y de las constantes q_k elegidas. Si se trabaja con una única variable auxiliar $X = X_1$, y $q_k = \frac{1}{x_k}$ y se realiza un muestreo aleatorio simple, entonces el total se calcula como un estimador de razón:

$$\hat{T}_{Y_{reg}} = \frac{\overline{x}}{\overline{y}} \cdot X \quad [3.24]$$

donde $\overline{y}$ y $\overline{x}$ son las correspondientes medias muestrales de la variable Y y la variable X. El estimador $\hat{T}_{Y_{reg}}$ no es generalmente insesgado pero los pesos ω_k serán muy próximos a d_k, por lo que es asintóticamente insesgado (Särndal, 2007).

Existen alternativas para la construcción de los estimadores de calibración que no modifican la medida de distancia, sino que cambian el proceso de construcción del estimador a través de la modificación de alguna de estas dos condiciones (y en muchos casos, apoyándose en modelos de superpoblación):

- Minimización de una distancia.
- Que los pesos equilibrados den estimaciones perfectas para las variables auxiliares.

Martínez (2002) menciona cuatro tipos de estimadores de calibración, entre los que el más flexibles es el asistido por modelos:

- Estimadores de calibración para una familia de distancias.
- Estimadores de calibración basados en una forma funcional.
- Estimadores de calibración cosméticos.
- Estimador de calibración asistido por modelos.

Wu (2001) clasifica los métodos asistidos por en tres tipos:

- Estimadores de regresión generalizada (GREG) (Cassel *et al.*, 1976) y (Särndal, 1980).
- Estimadores de calibración (Deville y Särndal, 1992).
- Empíricos de probabilidad (Chen y Qin, 1993; Chen y Sitter, 1999).

Los estimadores de calibración asistidos por modelos se construyen al sustituir en el proceso de calibración la restricción de la expresión [3.25] por otra más adecuada, ya que se asume implícitamente un modelo lineal de regresión, entre la variable de estudio Y y las variables del vector X, en la población de estudio:

$$\sum_{k \in S} \omega_k \cdot y_k = T_X \qquad [3.25]$$

La relación entre ambas variables no tiene porque acomodarse a una regresión lineal simple, de ahí la necesidad de plantear una mecanismo más flexible que tome una función de relación que se adapte a cada situación. Para ello se puede utilizar un modelo sobre un estimador de regresión generalizada, denominada GREG (Deville y Särndal, 1992; Särndal *et al.*, 2003). Este tipo de calibración, asume que la relación entre las variables Y y el vector X se describe por un modelo de superpoblación, especificado como:

$$\begin{aligned} E_\xi(y_k|X_k) &= \mu(X_k, \theta) \\ V_\xi(y_k|X_k) &= v_k^2 \cdot \sigma^2 \end{aligned} \qquad [3.26]$$

Con $k = 1, 2, ..., N$, donde $\theta = (\theta_0, \theta_1, ...\theta_J)'$ y σ^2 son parámetros poblacionales desconocidos, $\mu(X, \theta)$ es una función conocida de X y θ, v_k es una función conocida de X_k o bien de $\mu_k = \mu(X_k, \theta)$ y $E\xi$ y $E\xi$ son, respectivamente, la esperanza y la varianza con respecto al modelo de superpoblación.

Esta especificación general incluye tres de los casos más comunes de calibración con modelos que son: los modelos de regresión lineales, los no lineales y los generalizados.

Los enfoques anteriores se basan en un contexto de un modelo de regresión e incorporan esencialmente las variables auxiliares a través de sus medias poblacionales conocidas, incluso cuando se conocen las variables auxiliares para cada unidad en el población. Siendo $1/\pi_k$ los pesos ordinarios del muestreo para la observación k-ésima, dónde π_k es la probabilidad de inclusión de k:

$$V_\xi(y_k|X_k) = v(\mu_k)\ k = 1, 2, ...N \quad [3.27]$$

donde $\mu_k = E_\xi(y_k\ X_k)$ es una función de enlace y V es la función varianza. El estimador de calibración asistido por un modelo T_Y se define como:

$$\hat{T}_Y = \sum_{k \in S} \omega_k\ y_k \quad [3.28]$$

donde los pesos calibrados w_k son mínimos, según una medida de distancia con respecto a d_k, estando $\hat{T}_Y$ sujeto a la condición:

$$\sum_{k \in S} \omega_k \mu(X_k \hat{\theta}) = \sum_{k=1}^{N} \mu(X_k \hat{\theta}) \quad [3.29]$$

una vez minimizada la medida de distancia, se obtiene el siguiente estimador:

$$\hat{T}_Y^* = \hat{T}_{Y_\pi} + \left(\sum_{k=1}^{N} \hat{\mu_k}\ -\ \sum_{k \in S} d_k \hat{\mu_k} \right) \ddot{B}_N^* \quad [3.30]$$

donde:

$$\ddot{B}_N^* = \frac{\sum_{k \in S} d_k \cdot q_k \hat{\mu}_k \cdot y_k}{\sum_{k \in S} d_k \cdot q_k \hat{\mu}_k^2} \quad [3.31]$$

Wu (2001) identifica las propiedades más importantes del estimador de calibración:

1. Bajo ciertas condiciones, tanto $\hat{T}_Y$ como $\hat{T}_Y^*$ son iguales a $\hat{T}_{Y_\pi} + O\left(n^{-1/2}\right)$ y son asintóticamente insesgados para T_Y, con respecto al diseño, sin tener en cuenta si el modelo es correcto o no.
2. Si $q_k = 1/v_k^2$, entonces $\hat{T}_Y$ y $\hat{T}_Y^*$ pueden ser estimadores de calibración basados modelos, tanto de tipo lineal como no lineal. Es decir, ambos son consistentes respecto al diseño, independientemente del tipo de modelo. En el caso de un modelo sin error, es decir, $y_k = \mu_k$, la condición puede expresarse como:

$$\hat{T}_Y = \hat{T}_Y^* = \hat{T}_Y \quad [3.32]$$

Además, si se usa un modelo lineal ambos, $\hat{T}_Y$ y $\hat{T}_Y^*$, se reducen al estimador convencional de calibración (Deville y Särndal, 1992).

Si la encuesta parte de una muestra de tamaño n, donde w es el vector de pesos originales, de dimensión $n \times 1$, y ω' el vector homólogo de pesos transformados, cualquier procedimiento de reponderación que se aplique dará lugar a una relación funcional del tipo $\omega' = \omega'(\omega, X)$. De manera que los nuevos pesos van a ser función de los originales y de las variables auxiliares elegidas.

La información auxiliar proporcionada por la encuesta va a estar contenida en una matriz $X_{n \cdot p}$ donde en cada fila aparecen los valores de las variables auxiliares para cada individuo de la muestra. Los nuevos pesos han de cumplir la condición de equilibrado de la muestra, es decir, $X'\omega' = x$, siendo x el vector de efectivos poblacionales proporcionados por las fuentes externas utilizadas. Con los pesos ω', se calcularían las nuevas estimaciones para cualquier variable Y de interés.

Medidas de distancia en la calibración

La calibración parte de la definición previa de una función de distancia $G(\omega, \omega')$ entre los pesos originales ω y los calibrados ω', para la que se exige que:

$$\sum_{k=1}^{n} \omega_k \cdot G(\omega_k \cdot \omega'_k) = N \qquad [3.33]$$

sea mínimo para el conjunto de la muestra, con la restricción:

$$\sum_{k=1}^{n} \omega'_k = N \qquad [3.34]$$

Es decir, que la suma de pesos transformados debe recuperar un determinado total de población. Siendo h el cociente entre ponderaciones w'_k/w_k, se definen las dos familias de distancias más usualmente utilizadas:

- Cuadrática: $G(h) = \left(\frac{h-1}{2}\right)^2$.
- Logarítmica $G(h) = h \;\; log(h) - h + 1, h > 0$.

Tabla 3.16. Medidas de distancia básicas para calibración

Medida de distancia	Especificación
Chi cuadrado	$(\omega - d)^2/2qd$
Chi cuadrado modificado	$(\omega - d)^2/2q\omega$
Mínima entropía	$q^{-1}(-d\log(\omega/d) + \omega - d)$
Mínima entropía modificada	$q^{-1}(w\log(\omega/d) - \omega - d)$
Hellinger	$2(\sqrt{\omega} - \sqrt{d})^2/q$

Existen múltiples forma de especificar la medida de distancia $D(\omega, d)$. Deville y Särndal (1992) recogen 5 medidas de distancia, mostradas en la Tabla 3.16.

Es importante señalar que, dependiendo de la función de distancia elegida $D(\omega, d)$, puede que no exista una solución analítica para la condición:

$$\frac{\partial Q}{\partial \omega_i} = \frac{(\omega_i - d_i)}{q_i d_i} - \lambda x_i \qquad [3.35]$$

y es posible que requiera una aproximación numérica de w_i usando Newton-Raphson o un método similar. Además, la solución a la ecuación [3.35] puede producir ponderaciones positivas, negativas o extremadamente grandes que pueden ser no ser deseables en un contexto de muestreo.

En términos de eficiencia, Deville y Särndal (1992) demostraron que para muestras de tamaño medio y grande, la elección de la función $D(\omega, d)$ no tiene un gran impacto en la varianza del estimador elegido. Deville y Särndal también demostraron que bajo ciertas condiciones, el estimador es asintóticamente equivalente a GREG para cualquier función de distancia $D(w, d)$. Por lo tanto, la elección de la función de distancia no es importante para muestras grandes, sino que depende del esfuerzo de proceso de resolver la condición de la expresión [3.35].

Métodos de calibración

Asociados a estas funciones de distancia, existen diversos métodos de calibración, que proponen funciones de transformación de los pesos nuevos respecto a los originales. Entre los más comunes, se encuentran los métodos lineal, exponencial y lineal *logit* truncado.

La transformación lineal se basa en la medida de distancia Chi-cuadrado, y es la más comúnmente aplicada por su sencillez y porque ofrece generalmente buenos resultados, la medida de distancia se definiría como:

$$\omega' = \omega \cdot (1 + u) \qquad [3.36]$$

La transformación lineal tiene dos efectos no deseados: el primero, que puede obtener pesos negativos; y el segundo, que ofrece valores no acotados, y por tanto es probable que los pesos puedan tomar valores muy extremos.

Por otra parte, existe el método exponencial que se basa en la medida de distancia:

$$\omega_k \cdot log\left(\frac{w_k}{d_k}\right) - w_k + d_k \qquad [3.37]$$

En este caso, siempre se generan pesos positivos, pero en cambio, puede haber una mayor distorsión de pesos nuevos respecto a los originales. Para este método, como para el siguiente *logit*, es recomendable establecer cotas a la transformación de los pesos originales, es decir, se buscan dos valores L y U tal que $L < h_k < U, k = 1, 2....n$ donde $h_k = w'_k/w_k$.

Los métodos *logit* lineal y truncado, también son métodos acotados, lo que significa que ofrecen límites superior e inferior sobre las relaciones de peso $\left(\frac{w_k}{d_k}\right)$, que se denominan pesos g y que permiten controlar las transformaciones de pesos extremos. Estas relaciones no son arbitrarias y dependen de las variables de calibración elegidas, por lo general, para determinar las cotas se realiza un proceso iterativo de descubrimiento de los límites. Como propone Rao (1996), se establece inicialmente un intervalo amplio para h_k que se va reduciendo progresivamente mientras se consiga una solución.

El procedimiento es fácilmente generalizable a múltiples dimensiones. Así, si suponemos 3 dimensiones para el caso lineal, por ejemplo, provincia, grupo de sexo y edad, los nuevos pesos se calcularían según:

$$\omega'_k = \omega_k \cdot (1 + x_{prov} + y_{sexo} + z_{edad}) \qquad [3.38]$$

donde las cantidades x_{prov}, y_{sexo} y z_{edad} serían las incógnitas a resolver, que aunque pueden tomar cualquier signo, normalmente al sumarse ofrecen una magnitud cercana a cero, para así satisfacer mejor la condición de variación mínima en la transformación de los pesos.

Elección de datos auxiliares para la calibración

La elección de las fuentes de información para la calibración no es sencilla, se asume que tanto las variables y sus totales son completos y exactos, aunque en la práctica no suele cumplirse. En los casos más extremos, los errores o distorsiones en la información auxiliar pueden llegar a dañar seriamente los pesos calibrados.

Los métodos de calibración permiten mejorar la estimación de parámetros mediante la incorporación de fuentes adicionales[45], mitigan los problemas de falta de respuesta, mejoran el ajuste de los pesos poblacionales y permiten

[45] Por ejemplo, se pueden tomar los totales de calibración de los resultados de otra encuesta, como hace la Oficina Central de Estadísticas de Irlanda para las estimaciones de su encuesta de población activa, que usa para calibrar los datos de las estadísticas de Eurotat de ingresos y condiciones de vida (EU-SILC). Esta última fuente incorpora, entre otros, nivel de ingresos, tasa de pobreza, inclusión social o pensiones.

obtener estimaciones consistentes entre las variables auxiliares y la muestra. Pero para que el proceso sea válido, es necesario asegurar la consistencia y calidad de la información de las variables auxiliares, por tanto debe asegurase que:

- Las estimaciones usadas como variables auxiliares sean (casi) imparciales y provengan de una muestra que sea, como mínimo, del mismo tamaño.
- Que ambas encuestas sean consistentes en cuanto a las magnitudes que miden, por ejemplo si ambas manejan ingresos, estas debe referirse a la misma magnitud.

Evaluación de la calidad del proceso

El control de la validez de todos pasos del proceso desde control de calidad y coherencia de las fuentes, pasando por las técnicas de modelado y terminando por los algoritmos de imputación, tienen un enorme impacto en la calidad de los resultados. Aún cuando se aseguren ciertos niveles requisitos de coherencia en los datos de entrada, los resultados se deben validar en función de su capacidad de proporcionar estimaciones fiables y precisas.

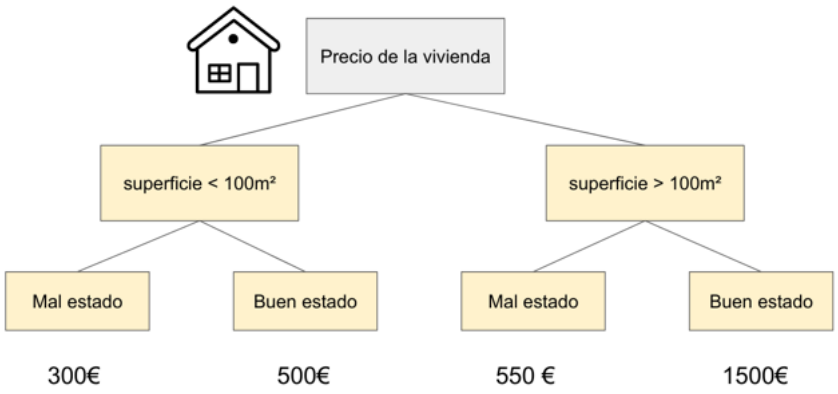
Rässler (2012) propone un marco de trabajo para la evaluación de la calidad a través de cuatro niveles de validación: 1) las distribuciones marginales y conjuntas de las variables del fichero donante se preservan en el fichero final; 2) la estructura de la correlación y los momentos altos de las variables se preservan después de la correspondencia estadística; 3) la distribución verdadera conjunta de todas las variables se refleja en el fichero final; y 4) los valores verdaderos pero desconocidos de la variable Z del fichero receptor se reproducen.

En el proceso es esencial tener en cuenta la incertidumbre, y en particular aquella asociada a las asunciones *a priori* implícitas en el modelo. Ante las limitaciones metodológicas, se proponen dos enfoques (D'Orazio *et al.*, 2006): 1) estimar la incertidumbre en las estimaciones finales, lo que está generalmente enfocado a macro-objetivos (estimación de coeficientes de correlación y tablas de contingencia); 2) Centrado en la identificación de información auxiliar que permita reducir la incertidumbre y pueda relajar las condiciones de independencia condicional.

Anexo 3b. Algoritmo Random Forests

Los modelos de tipo Random Forests, o bosques aleatorios, son un tipo de modelo de aprendizaje estadístico que usan árboles de decisión o regresión. Para la valoración de la vivienda, un modelo de regresión basado en árboles, estima el precio en base a una serie de "reglas de decisión". En el ejemplo gráfico de la Figura 3.28 se muestra cómo un modelo de tipo árbol estima el precio de una vivienda, para llegar al precio final se siguen una serie de reglas que acotan el precio en base a sus características.

Figura 3.28. odelo de valoración basado en un árbol de decisión simple



Fuente: elaboración propia.

A partir de los datos de entrada, las reglas de decisión (cortes) del árbol se calculan en función de su capacidad para reducir de la entropía en los grupos formados después del corte. Este desorden se puede expresar como impureza de los subárboles generados, ganancia/pérdida de información, coeficiente Gini o la varianza. Los modelos de árbol de regresión son una alternativa eficaz al análisis de regresión múltiple (Breiman, 2017; Fan *et al.*, 2006).

En conjuntos grandes y con una gran cantidad de variables, los modelos de árboles simples pueden incurrir en problemas de infrajuste o sobreajuste. Para resolver estos problemas, se aplican los enfoques basados en ensamblados de árboles como los basados en *bagging*[46], *boosting*[47] o *stacking*[48].

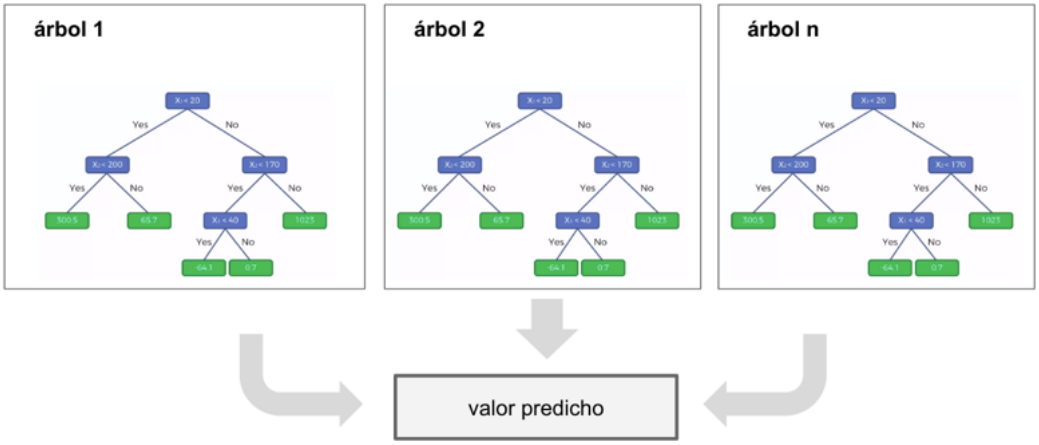
La técnica de Random Forests (Breiman, 2001), es un modelo de árbol de tipo ensamblado y fue desarrollada originalmente por Leo Breiman y Adele Cutler. Combina la idea de *bagging* y la selección aleatoria de atributos para construir

[46] El *bagging* consiste en combinar modelos distintos en paralelo con el objetivo de reducir la varianza.
[47] El *boosting* construye una secuencia de modelos predictivos orientados a corregir los errores del modelo anterior, para mejorar la precisión general del modelo final.
[48] El *stacking* utiliza las predicciones de varios modelos base para entrenar un modelo objetivo que tiene un mejor rendimiento predictivo que los originales.

una colección de árboles de decisión. Este algoritmo puede utilizarse para estimar valores categóricos (binarios o multiclase), denominado árbol de clasificación, o también para estimar una magnitud continua, dónde se lo denomina como árbol de regresión.

El modelo de *bagging* funciona construyendo una multitud de árboles de decisión en el momento del entrenamiento y generando una predicción media (regresión) de los árboles individuales (Figura 3.29). Este enfoque resuelve la tendencia al sobreajuste de otras modalidades de árboles ensamblados[49].

Figura 3.29. Esquema general de un modelo basado en Random Forests



Fuente: elaboración propia.

La idea esencial del *bagging* es promediar muchos modelos ruidosos pero aproximadamente insesgados (Hastie *et al.*, 2017), para producir un modelo combinado capaz de reducir la varianza. Si bien este proceso no fuerza el uso de un tipo de modelo en concreto, los modelos de árbol son los candidatos ideales para el *bagging*, dado que pueden registrar estructuras de interacción complejas en los datos, y si crecen con suficiente profundidad, tienen relativamente bajo sesgo. Dado que los árboles son notoriamente ruidosos, se benefician enormemente de la estimación basada en el promedio.

Cada árbol se construye a través de los siguientes pasos:

- Sea N el número de casos de prueba, y M es el número de variables en el clasificador.

[49] El sobreajuste se refiere la incapacidad de generalizar de un modelo, que no obstante muestra bajos niveles de error sobre el conjunto de datos de entrenamiento. Por tanto el modelo se ha especializado en replicar el dato de entrenamiento no en crear reglas que permitan evaluar correctamente otros datos.

- Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado, m debe ser un número mucho menor que M.
- Se elige un conjunto de entrenamiento para este árbol y se usa el resto de los casos de prueba para la estimación del error.
- Para cada nodo del árbol, se eligen aleatoriamente m variables, en las cuales basar la decisión y se calculan las mejores particiones del conjunto de entrenamiento, a partir de las m variables.

Finalmente, la inferencia se realiza recorriendo descendente cada árbol de decisión, tomando el valor del nodo terminal al que se llega a partir de las variables de entrada. Este proceso se realiza sobre todos los árboles en el ensamblado, y la estimación final se calcula como el promedio de las estimaciones individuales.

Antipov y Pokryshevskaya (2012), en su análisis sobre el modelo de *Random Forests* aplicado al ámbito inmobiliario, argumenta que esta técnica es una de las más adecuadas para la valoración de la vivienda por varios motivos:

- Muestra buenos resultados comparado con otras técnicas, como las máquinas de soporte vectorial (SVM), redes neuronales u otro tipo de modelos de árboles complejos como el *boosting*.
- Maneja satisfactoriamente variables categóricas con un gran número de niveles, sin incrementar el número de parámetros (como sería el caso de las redes neuronales que requieren la creación de variables ficticias *dummies*) lo que reduce la posibilidad del sobreajuste al introducir un gran volumen de variables dicotómicas.
- Funciona correctamente cuando existen valores ausentes, y no requiere procesos de imputación ni la eliminación de observaciones por este motivo.
- El proceso de *bagging* hace que el modelo sea robusto ante valores atípicos, ya que aparecerán menos en el muestreo de creación de árboles individuales (muestras de *bootstrap*), y por tanto su influencia en los resultados se ve reducida.
- Al contrario de los modelos de árboles de regresión simple (como CART), la estimación es un único valor, no una serie de valores discretos derivados de una serie de reglas.
- Los árboles permiten la gestión de las no linealidades, la heterocedasticidad y los comportamientos diferenciados las variables para los distintos segmentos, que los modelos lineales multivariantes.
- No se requiere una especificación detallada *a priori*.
- Las predicciones se encuentran en los mismos rangos que las observadas, lo que reduce la posibilidad de sobrestimación de las viviendas.

- Es posible medir la importancia de los factores, a través de su capacidad de reducción marginal de los errores, por parte de cada variable explicatoria.

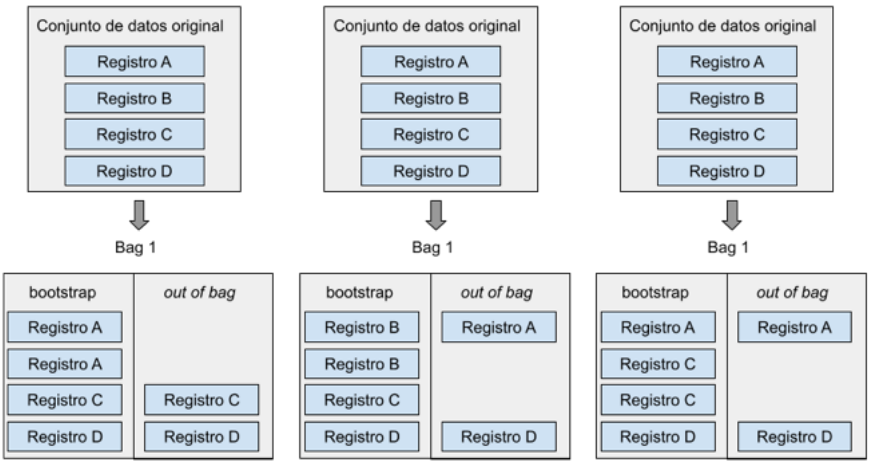
Muestra out-of-bag (OOB)

Una de las ventajas del método *Random Forests* es que puede generar métricas de error y de ajuste sin necesidad de un proceso de remuestreo previo que requiera dividir la muestra entre entrenamiento y validación. Para cada árbol construido, el algoritmo divide la muestra en dos conjuntos disjuntos: la *in bag* que se utiliza para construir el árbol de decisión y la *out of bag*, que se utiliza para calcular métricas sobre el modelo.

El error OOB, también es un método para medir el error de la predicción de *Random Forests*, aunque también es posible aplicarlo a otros tipos árboles de decisión basados *boosting*, y otros algoritmos de aprendizaje automático que utilizan agregación de tipo *bootstrap*[50](Hastie *et al.*, 2017).

Para comprender mejor la cuestión, en la Figura 3.30 se describe como se crea el conjunto OOB en el proceso de bagging. Primero se divide el conjunto en registros para entrenar (*in bag*) y para evaluar (*out of bag*) que no se utilizan en el entrenamiento, pero si para el cálculo de métricas de ajuste y error. Se realiza un proceso de sobremuestreo para completar el número de instancias de la muestra *"in bag"*. Este proceso se repite para cada árbol creado, de manera que las métricas del modelo se calculan promediando las medidas individuales.

Figura 3.30. Proceso de bagging mediante muestreo con reemplazo



Fuente: elaboración propia.

Con un suficiente número de árboles, las métricas OOB y las producidas la técnica

[50]Los métodos *bootstrap*,consisten en crear múltiples muestras de entrenamiento aleatorias con reemplazo de un conjunto de datos, para luego construir modelos a partir de estas muestras y combinar sus predicciones para mejorar la estabilidad y la precisión del modelo final.

de remuestreo denominadas de validación cruzada (LeCun *et al.*, 2015) producen una estimación similar.

Anexo 3c. Modelos GAM

Los modelos lineales tradicionales son simples, pero habitualmente adolecen de problemas cuando se aplican a situaciones reales, puesto que, los efectos rara vez suelen ser lineales (Hastie *et al.*, 2017).

La relación no lineal entre los predictores y la variable objetivo se puede controlar a través de flexibilizar los coeficientes de regresión, haciéndolos que sean función de las covariables, en lugar de constantes (Hastie y Tibshirani, 2017). Las funciones sobre las que se construyen estos coeficientes, de tipo funcional, se denominan funciones base, y son elemento central de los modelos aditivos generalizados.

En una regresión, un modelo aditivo generalizado tendría la forma siguiente:

$$E(Y|X_1, X_2, ..., X_p) = \alpha + f_1(X_1) + f_1(X_2) + ... + f_1(X_p) \quad [3.39]$$

donde $X_1, X_2, ..., X_p$ son los predictores, e Y la variable respuesta; las f_j son funciones de suavizado no paramétricas. Cada una de ellas se construye como una expansión de funciones base[51], para dar lugar a un modelo que se ajusta como una regresión simple de mínimos cuadrados.

En general, la medida condicional $\mu(X)$ de la variable de respuesta Y está relacionada con la función aditiva de los predictores mediante la función de enlace g:

$$g[\mu(X)] = \alpha + f_1(X_1) + f_1(X_2) + ... + f_1(X_p) \quad [3.40]$$

Ejemplo de función típicas son:

- $g(\mu)$ es el enlace identidad, se utiliza para mdoelos aditivos lineales con una respuesta gaussiana.
- $g(\mu) = logit(\mu)$ o $g(\mu) = probit(\mu)$ se utiliza para modelar probabilidades binomiales.
- $g(\mu) = log(\mu)$ se corresponde a modelos log-lineales o logarítmico-aditivos asociados a funciones de distribución de Poisson (conteos).

[51] La expansión se refiere a la suma de los resultados de las distintas funciones de suavizado

Los modelos aditivos se pueden entender como forma una extensión sobre los modelos lineales, que los hace más flexibles mientras que se mantiene su interpretabilidad. Sin embargo estos modelos pueden tener limitaciones en análisis con grandes volúmenes de datos y alta dimensionalidad (Hastie *et al.*, 2017), principalmente porque el método intenta estimar una función de suavizado para cada predictor. Se han planteado diversas aproximacion con el objeto de resolver los inconvenientes anteriores, como por ejemplo utilizar penalizaciones de tipo Lasso, denominadas COSSO (Lin y Zhang, 2006), o el método SpAM[52] (Ravikumar *et al.*, 2007).

Para grandes volúmenes de datos, también se puede aplicar métodos como el *boosting* para estimar la expansión de las funciones base. En los últimos años, se han publicado una serie de algoritmos que permiten estimar modelos GAM de forma eficiente en conjuntos grandes (Li y Wood, 2020; Wood *et al.*, 2015, 2017).

En el ámbito de la valoración de la vivienda existen diversas aplicaciones como el modelo desarrollado por Pace (1998) que intenta limitar el efecto de las no linealidades con un hedónico basado en GAM. A modo de ejemplo se puede señalar la aplicación al mercado de la vivienda en Alemania de Munger (2021), para Eslovenia encontramos a (Ulbl *et al.*, 2021), para Sudáfrica usando un método GAM jerárquico (Bax *et al.*, 2021).

[52] *Sparse Additive Models*.

Anexo 3d. Sumas calibración de la EPF

Las Tabla 3.17 muestra las sumas originales del conjunto de datos de la EPF, sobre las que se ha aplicado un proceso de suavizado exponencial para calcular las sumas de la calibración definitiva, y que se recogen en la Tabla 3.18.

Tabla 3.17. Totales originales para calibración EPF

Variable	2011	2012	2013	2014	2015	2016	2017	2018	2019
Total	428667	96%	95%	95%	99%	111%	113%	109%	122%
1 o 2 habitaciones	267959	106%	104%	106%	110%	120%	112%	118%	138%
3 habitaciones	126257	87%	78%	81%	84%	101%	115%	98%	99%
4 habitaciones	25964	89%	158%	82%	132%	144%	161%	131%	142%
5 o más habitaciones	8487	48%	39%	70%	39%	33%	40%	41%	68%
Menos de 60	174119	110%	93%	104%	108%	120%	105%	108%	122%
De 61 a 75	99464	77%	86%	97%	81%	89%	92%	93%	106%
De 76 a 90	86505	100%	93%	77%	117%	148%	155%	138%	174%
Más de 90	68579	98%	133%	95%	99%	109%	145%	119%	124%
Menos de 10 viviendas	81814	95%	114%	107%	99%	112%	130%	115%	113%
10 o más viviendas	346854	97%	90%	92%	99%	111%	109%	107%	124%
Chalé	13337	96%	173%	200%	210%	120%	158%	136%	154%
Casa media	362376	95%	92%	91%	102%	117%	120%	114%	127%
Casa económica	52954	104%	90%	86%	49%	68%	57%	65%	77%
Urbana alta	40882	97%	123%	71%	84%	84%	92%	61%	98%
Urbana media	365154	97%	95%	100%	102%	118%	119%	114%	123%
Urbana inferior	22631	81%	36%	46%	77%	51%	48%	110%	150%
Renta baja	188857	96%	92%	93%	87%	85%	79%	73%	54%
Renta media	186721	86%	95%	87%	110%	121%	127%	107%	130%
Renta alta	53089	117%	134%	149%	100%	186%	193%	213%	269%
Gasto bajo-medio	28825	112%	96%	163%	135%	187%	166%	128%	134%
Gasto medio-alto	170787	78%	67%	85%	105%	124%	128%	109%	112%
Gasto alto	229056	100%	104%	85%	90%	92%	98%	105%	123%
Ciudad de Madrid	264129	88%	83%	80%	88%	106%	98%	97%	103%
Resto CAM	164538	118%	127%	135%	129%	124%	152%	141%	172%
Zona densamente poblada	387080	95%	88%	86%	95%	107%	104%	100%	111%
Zona intermedia	26142	120%	172%	160%	128%	125%	169%	175%	213%
Zona diseminada	15446	97%	133%	192%	155%	196%	235%	215%	228%

Fuente: elaboración propia

Tabla 3.18. Totales suavizados exponencialmente para calibración EPF

Variable	2011	2012	2013	2014	2015	2016	2017	2018	2019
Total	428667	103%	102%	101%	100%	102%	109%	114%	114%
1 o 2 habitaciones	267959	105%	110%	112%	114%	117%	123%	124%	126%
3 habitaciones	126257	100%	93%	85%	83%	83%	92%	103%	100%
4 habitaciones	25964	105%	110%	116%	121%	126%	132%	137%	142%
5 o más habitaciones	8487	96%	68%	49%	55%	42%	33%	32%	32%
Menos de 60	174119	103%	106%	108%	110%	112%	115%	117%	118%
De 61 a 75	99464	101%	90%	89%	93%	88%	89%	91%	93%
De 76 a 90	86505	109%	116%	119%	117%	126%	141%	154%	159%
Más de 90	68579	103%	106%	109%	112%	115%	118%	121%	124%
Menos de 10 viviendas	81814	102%	103%	105%	107%	108%	110%	112%	114%
10 o más viviendas	346854	103%	103%	99%	98%	102%	109%	112%	112%
Chalé	13337	107%	108%	148%	181%	203%	168%	170%	160%
Casa media	362376	103%	103%	101%	99%	104%	114%	120%	121%
Casa económica	52954	97%	98%	91%	86%	64%	63%	57%	58%
Urbana alta	40882	100%	99%	102%	98%	96%	93%	91%	85%
Urbana media	365154	103%	103%	102%	104%	106%	114%	119%	120%
Urbana inferior	22631	106%	100%	73%	65%	77%	70%	64%	93%
Renta baja	188857	94%	89%	85%	83%	80%	77%	72%	67%
Renta media	186721	104%	105%	107%	107%	111%	117%	122%	123%
Renta alta	53089	121%	141%	160%	178%	175%	199%	218%	238%
Gasto bajo-medio	28825	104%	111%	111%	129%	135%	152%	160%	157%
Gasto medio-alto	170787	102%	91%	80%	84%	96%	112%	121%	117%
Gasto alto	229056	103%	104%	107%	99%	97%	97%	100%	105%
Ciudad de Madrid	264129	100%	95%	89%	85%	87%	97%	98%	98%
Resto CAM	164538	109%	118%	127%	136%	145%	154%	163%	172%
Zona densamente poblada	387080	101%	100%	95%	92%	95%	102%	105%	104%
Zona intermedia	26142	114%	128%	143%	157%	171%	185%	199%	213%
Zona diseminada	15446	116%	131%	147%	166%	181%	198%	216%	232%

Fuente: elaboración propia

Anexo 3e. Descriptivos de modelos hedónicos

Las Tablas 3.19 y 3.20 recogen los coeficientes de la regresión del mercado de la EPF, y las Tabla 3.23 los de regresión del modelo de oferta. Los coeficientes del modelo de conversión se muestran en la Tabla 3.25. En todos los casos anteriores, los términos de las funciones de suavizado de los modelos son altamente significativos, véanse las Tablas 3.22, 3.21 y 3.24 del presente anexo.

Tabla 3.19. Coeficientes regresión GAM - Modelo alquiler sobre EPF 2012 - 1/2

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	3.89	0.05	78.82	0.00	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	0.01	0.02	0.22	0.83	
TAMAMUMunicipio con 20.000 o más y menos de 50.000	-0.01	0.03	-0.25	0.80	
TAMAMUMunicipio con 10.000 o más y menos de 20.000	-0.09	0.03	-2.77	0.01	**
TAMAMUMunicipio con menos de 10.000 habitantes	-0.08	0.03	-2.56	0.01	*
TIPOEDIFVivienda unifamiliar adosada o pareada	-0.05	0.01	-5.14	0.00	***
TIPOEDIFCon menos de 10 viviendas	-0.07	0.02	-3.86	0.00	***
TIPOEDIFCon 10 ó más viviendas	-0.05	0.02	-3.00	0.00	**
TIPOEDIFOtros (destinado a otros fines o alojamien	-0.04	0.11	-0.38	0.70	
ZONARESUrbana alta	0.08	0.04	2.30	0.02	*
ZONARESUrbana media	0.02	0.03	0.43	0.66	
ZONARESUrbana inferior	0.02	0.04	0.50	0.62	
ZONARESRural industrial	-0.11	0.04	-2.76	0.01	**
ZONARESRural pesquera	-0.03	0.06	-0.56	0.57	
ZONARESRural agraria	-0.03	0.04	-0.84	0.40	
ANNOCONHace 25 ó más años	-0.01	0.01	-1.59	0.11	
DENSIZona intermedia	0.02	0.03	0.90	0.37	
DENSIZona diseminada	-0.05	0.03	-1.89	0.06	.
INTERINPSPDe 500 a menos de 1000 €	-0.01	0.02	-0.47	0.64	
INTERINPSPDe 1000 a menos de 1500 €	0.02	0.02	0.97	0.33	
INTERINPSPDe 1500 a menos de 2000 €	0.02	0.02	1.08	0.28	
INTERINPSPDe 2000 a menos de 2500 €	0.06	0.02	2.32	0.02	*
INTERINPSPDe 2500 a menos de 3000 €	0.03	0.03	1.26	0.21	
INTERINPSP3000 o más €	0.02	0.03	0.80	0.43	
NHABIT3 habitaciones	-0.03	0.02	-1.71	0.09	.
NHABIT4 habitaciones	-0.05	0.02	-2.67	0.01	**
NHABIT5 o más habitaciones	-0.05	0.02	-2.50	0.01	*
CAPROVNo	-0.10	0.02	-5.71	0.00	***
factorGASTOT6De.15.83.a.16.26	0.02	0.02	0.73	0.46	
factorGASTOT6De.16.26.a.16.62	0.04	0.02	1.75	0.08	.
factorGASTOT6De.16.62.a.17	0.02	0.02	0.84	0.40	
factorGASTOT6De.17.a.17.46	0.05	0.02	2.25	0.02	*
factorGASTOT6Más.de.17.46	0.09	0.02	3.75	0.00	***

Fuente: elaboración propia

Tabla 3.20. Coeficientes regresión GAM - Modelo alquiler sobre EPF 2012 - 2/2

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
CCAAAragón	0.09	0.03	2.94	0.00	**
CCAAAsturias, Principado de	0.17	0.04	4.76	0.00	***
CCAABalears, Illes	0.30	0.04	7.49	0.00	***
CCAACanarias	0.06	0.02	2.77	0.01	**
CCAACantabria	0.28	0.03	9.75	0.00	***
CCAACastilla y León	-0.01	0.02	-0.46	0.64	
CCAACastilla-La Mancha	0.05	0.02	2.52	0.01	*
CCAACataluña	0.29	0.02	17.87	0.00	***
CCAAComunitat Valenciana	-0.02	0.01	-1.24	0.22	
CCAAExtremadura	-0.10	0.03	-2.95	0.00	**
CCAAGalicia	0.04	0.02	1.95	0.05	.
CCAAMadrid, Comunidad de	0.22	0.02	11.59	0.00	***
CCAAMurcia, Región de	-0.05	0.02	-2.56	0.01	*
CCAANavarra, Comunidad Foral de	0.43	0.03	12.23	0.00	***
CCAAPaís Vasco	0.50	0.06	8.12	0.00	***
CCAARioja, La	0.23	0.09	2.64	0.01	**
CCAACeuta	0.36	0.09	3.95	0.00	***
CCAAMelilla	0.07	0.09	0.74	0.46	

Fuente: elaboración propia

Modelo de alquiler EPF para 2012

Tabla 3.21. Términos de la función de suavizado - Modelo alquiler sobre EPF 2012

Término	edf	Ref. df	F	p-value	signif.
s(SUPERF)	7.59	8.47	63.43	0	***

Fuente: elaboración propia

Modelo de idealista para 2012

Tabla 3.22. Términos de la función de suavizado - Modelo idealista 2012

Término	edf	Ref. df	F	p-value	signif.
s(SUPERF2)	8.74	8.98	384.83	0	***

Fuente: elaboración propia

Tabla 3.23. Coeficientes regresión GAM - Modelo oferta 2012

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	4.44	0.06	68.49	0.00	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	0.07	0.01	5.66	0.00	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	0.03	0.01	2.18	0.03	*
TAMAMUMunicipio con 10.000 o más y menos de 20.000	0.04	0.02	2.25	0.02	*
TAMAMUMunicipio con menos de 10.000 habitantes	-0.15	0.02	-9.19	0.00	***
TIPOEDIFVivienda unifamiliar adosada o pareada	0.26	0.01	29.96	0.00	***
ZONARESUrbana alta	-0.24	0.02	-11.65	0.00	***
ZONARESUrbana media	-0.25	0.03	-8.50	0.00	***
ZONARESUrbana inferior	-0.22	0.04	-5.35	0.00	***
ANNOCONHace 25 ó más años	0.01	0.01	2.40	0.02	*
DENSIZona intermedia	-0.05	0.01	-3.97	0.00	***
DENSIZona diseminada	-0.11	0.02	-7.22	0.00	***
INTERINPSPDe 500 a menos de 1000 €	0.03	0.01	1.73	0.08	.
INTERINPSPDe 1000 a menos de 1500 €	0.11	0.02	7.21	0.00	***
INTERINPSPDe 1500 a menos de 2000 €	0.25	0.02	14.54	0.00	***
INTERINPSPDe 2000 a menos de 2500 €	0.38	0.02	18.43	0.00	***
INTERINPSPDe 2500 a menos de 3000 €	0.46	0.03	17.05	0.00	***
INTERINPSP3000 o más €	0.47	0.03	15.01	0.00	***
NHABIT3 habitaciones	-0.05	0.01	-4.61	0.00	***
NHABIT4 habitaciones	-0.03	0.01	-2.63	0.01	**
NHABIT5 o más habitaciones	0.01	0.01	0.58	0.56	
CAPROVNo	-0.22	0.01	-17.73	0.00	***
factorGASTOT6De.16.26.a.16.62	-0.05	0.06	-0.72	0.47	
factorGASTOT6De.16.62.a.17	-0.04	0.06	-0.64	0.52	
factorGASTOT6De.17.a.17.46	-0.06	0.06	-1.00	0.32	
factorGASTOT6Más.de.17.46	-0.02	0.06	-0.39	0.69	

Fuente: elaboración propia

Modelo de correspondencia oferta-alquiler unifamiliar 2012

Tabla 3.24. Términos de la función de suavizado - Modelo conversion 2012

Término	edf	Ref. df	F	p-value	signif.
s(preciom2_anualpred)	8.99	9	1641.03	0	***
s(SUPERF2)	9.00	9	25917.82	0	***

Fuente: elaboración propia

Tabla 3.25. Coeficientes regresión GAM - Modelo conversion 2012

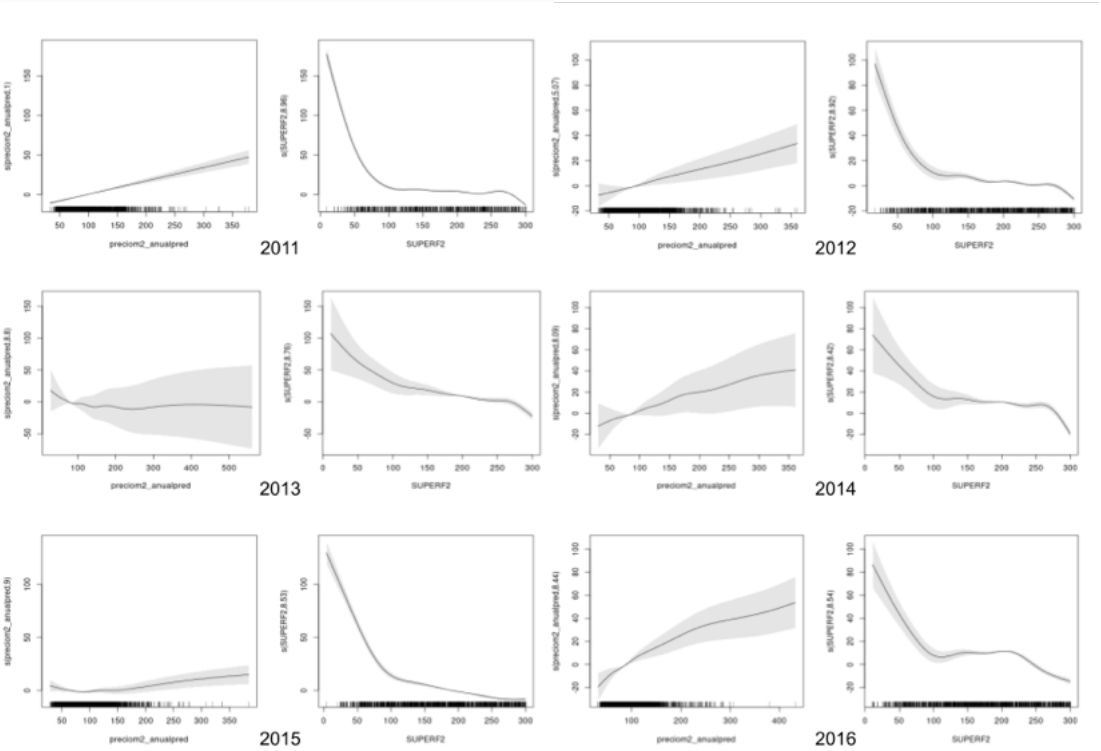
Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	85.78	0.55	156.76	0	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	6.97	0.09	77.16	0	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	8.23	0.14	59.88	0	***
TAMAMUMunicipio con 10.000 o más y menos de 20.000	7.18	0.19	37.44	0	***
TAMAMUMunicipio con menos de 10.000 habitantes	7.20	0.25	28.71	0	***
TIPOEDIFCon 10 ó más viviendas	1.65	0.07	23.46	0	***
TIPOCASACasa económica o alojamiento	-4.88	0.12	-40.46	0	***
ZONARESUrbana alta	-1.66	0.32	-5.14	0	***
ZONARESUrbana media	-1.71	0.38	-4.47	0	***
ZONARESUrbana inferior	-7.62	0.41	-18.46	0	***
ANNOCONHace 25 ó más años	0.84	0.05	15.79	0	***
DENSIZona intermedia	-6.42	0.14	-45.88	0	***
DENSIZona diseminada	-16.65	0.26	-63.47	0	***
INTERINPSPDe 500 a menos de 1000 €	0.65	0.12	5.44	0	***
INTERINPSPDe 1000 a menos de 1500 €	3.41	0.13	26.00	0	***
INTERINPSPDe 1500 a menos de 2000 €	7.97	0.17	47.63	0	***
INTERINPSPDe 2000 a menos de 2500 €	12.57	0.19	65.14	0	***
INTERINPSPDe 2500 a menos de 3000 €	11.46	0.29	38.98	0	***
INTERINPSP3000 o más €	8.30	0.35	23.56	0	***
NHABIT3 habitaciones	-0.95	0.06	-15.42	0	***
NHABIT4 habitaciones	0.65	0.12	5.39	0	***
NHABIT5 o más habitaciones	1.28	0.29	4.49	0	***
CAPROVNo	-11.86	0.09	-138.88	0	***
factorGASTOT6De.16.26.a.16.62	8.03	0.46	17.44	0	***
factorGASTOT6De.16.62.a.17	11.33	0.39	29.40	0	***
factorGASTOT6De.17.a.17.46	10.73	0.38	28.60	0	***
factorGASTOT6Más.de.17.46	14.08	0.38	37.26	0	***

Fuente: elaboración propia

Funciones de suavizado en modelo de correspondencia

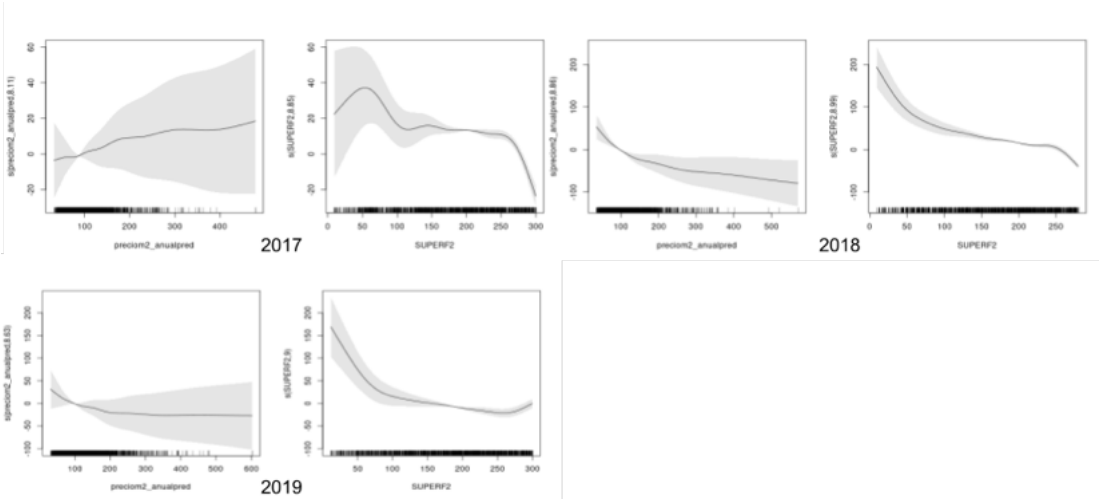
Las Figuras 3.31 y 3.32 muestran el comportamiento de las funciones de suavizado (*s*) de los modelos GAM de correspondencia, sobre el precio de oferta (preciom2_anualpred) y la superficie útil (SUPERF2).

Figura 3.31. Funciones de suavizado (s), vivienda unifamiliar: 2011-2016



Fuente: elaboración propia.

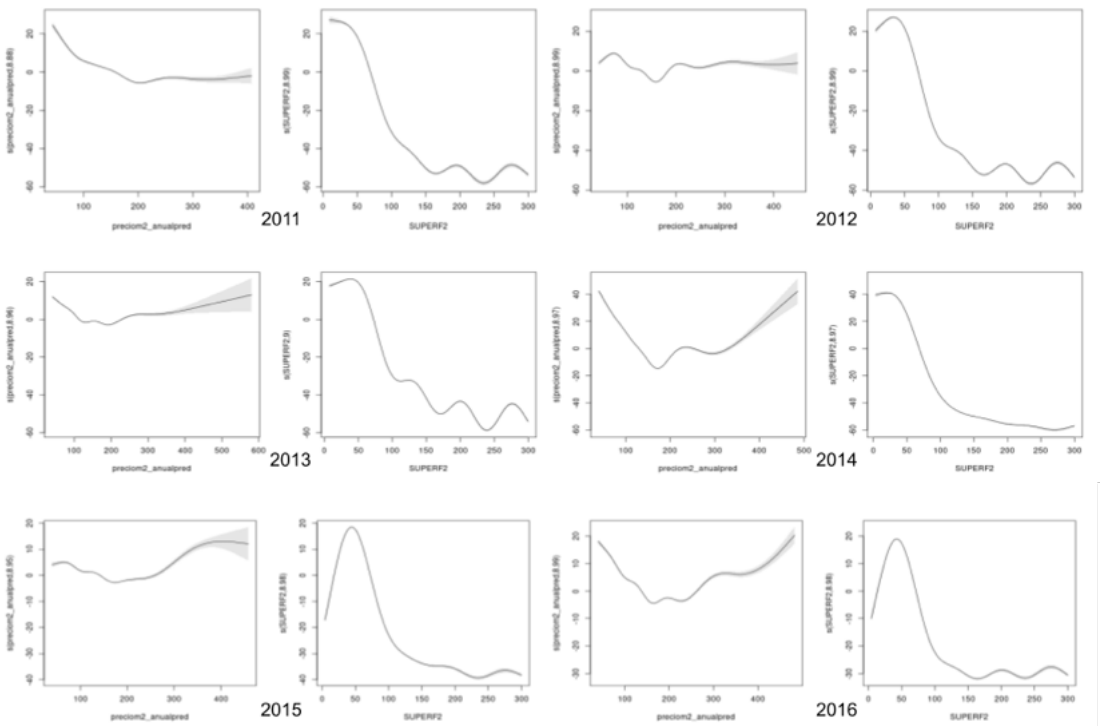
Figura 3.32. Funciones de suavizado (s), vivienda unifamiliar: 2017-2019



Fuente: elaboración propia.

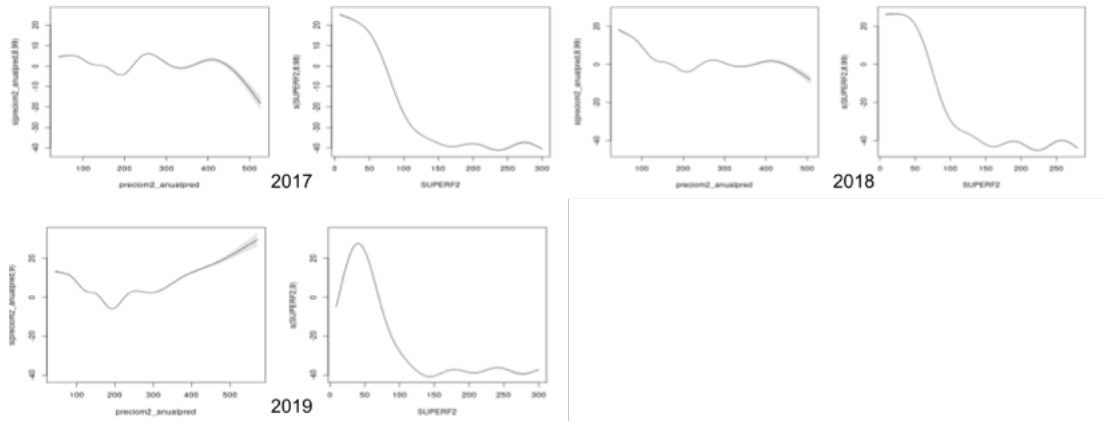
Las Figuras 3.33 y 3.34 muestran el comportamiento de las funciones de suavizado para las viviendas plurifamiliares. Se puede comprobar que las funciones de suavizado son mucho más significativas que en el caso de la vivienda plurifamiliar.

Figura 3.33. Funciones de suavizado (s), vivienda plurifamiliar: 2011-2016



Fuente: elaboración propia.

Figura 3.34. Funciones de suavizado (s), vivienda plurifamiliar:2017-2019



Fuente: elaboración propia.

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. Esa 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 espacial[1] y la dependencia espacial[2] (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

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

[2] En 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 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 viviendas[3]. 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 embardo, 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.* (Rey *et al.*).

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 gravitatorio[4] 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 dimensión[5], e inversamente proporcional a su distancia; el segundo, se centra en la distancia de un lugar a otro, ya sea la distancia euclidiana 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 espacial[6] (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 puja por lugares

[5] El 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.

[6] La 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 localización[7], 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

[7] La 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 euclidiana 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 OSM[8] 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, Ĉeh (2018) aplica la información de puntos de interés a la construcción de características de

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

esta naruraleza 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 anteriores*[9]*..."*. 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

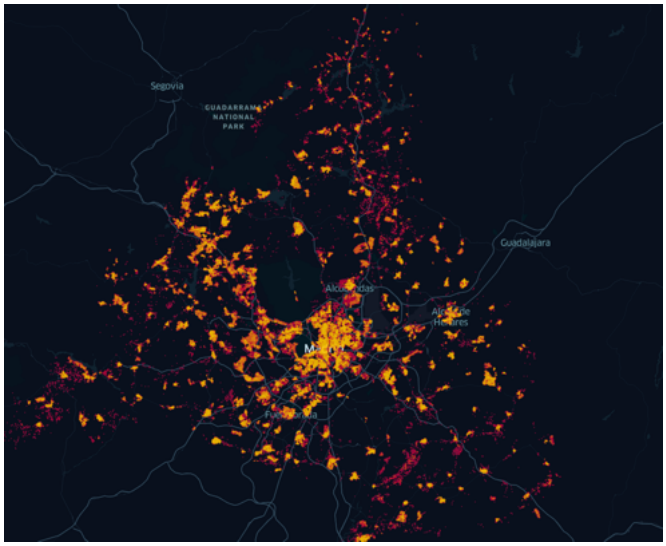
[9] *Referido a reproducir un proceso espacial observado para un conjunto de observaciones.*

anuncios de Idealista del año 2018, no solamente registros de alquiler sino que también de compraventa[10].

Las medidas de accesibilidad se construyen agregando superficies catrastrales, 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 centroide[11] 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.

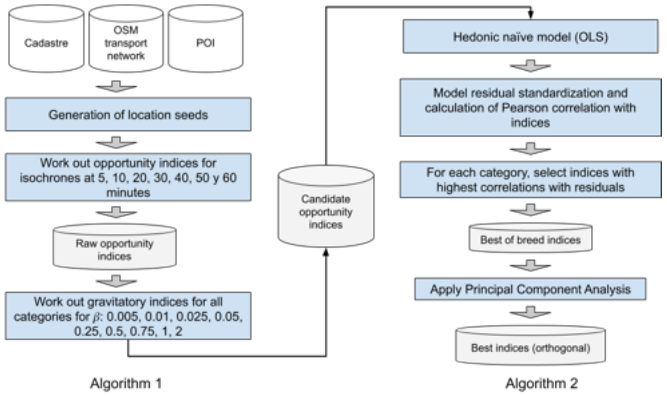
La metodología propuesta crea un conjunto de índices de accesibilidad que capturan la contribución de la ubicación a los precios de la vivienda, efecto ligado al principio de utilidad de la localización (Rey-Blanco *et al.*, 2023b). De una forma muy resumida, el proceso desarrollado para construir los 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

[10] Se opta por incluir datos de compraventa para incrementar el nivel de soporte del método.
[11] El 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 desplazamiento[12], 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

[12] En 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 $q = 1, ...Q$, alcanzable por coche o andando, *m*={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 $t = 1, ..., T$ períodos , más un término de error aleatorio ε_n^t. Es decir,

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

donde: $\hat{A}_{nk}^{tm*}$ representa los índices de *accesibilidad* (*A*) de ubicación en términos de $k = 1, ..., K$ *oportunidades* de ubicación a las que se puede llegar en automóvil y caminando; S_{nj}^t denota los atributos $j =, 1..., J$ *estructural* (*S*) de la propiedad; M_{nl}^t captura las características de oferta y demanda $l = 1, ..., L$ del sub*mercado* (*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 ˆ), 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 si, 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: $i(coche) = \{1, ..., I\} = \{5, 10, 20, 30, 40, 50\ y\ 60\ minutos\}$. Para los índices peatonales $d_{i(m=a\ pie)}$, se consideran los tiempos[14]. 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: $i(coche) = 1, ..., I = 5, 10, 20, 30, 40, 50 y 60 minutos$

[13] Un modelo sencillo sin especificación de variables de zona.

[14] La 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 $i(a\ pie) = \{1, ..., I\} = \{5, 10, 20\ y\ 30\ minutos\}$. 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:

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

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 $k = 1, ..., K$ 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
	aeropuerto	TRANSPORT.AIRPORT	frecuencia	OSM
Transporte privado	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
	salud	CAD.HEALTH	superficie	catastro
	comercio	SHOP	frecuencia	OSM
	comercio	CAD.COMMERCE	superficie	catastro
	agricultura	CAD.AGRICULTURE	superficie	catastro
	venues	CAD.VENUES	superficie	catastro
Social	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 usado 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.

[15] Es 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 (DGGS[16]), 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 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.

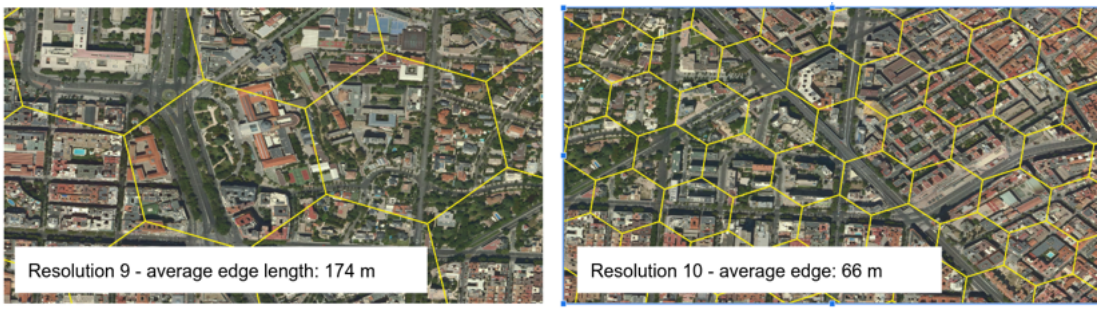
En el espacio de coordenadas teselado, los atributos de ubicación se precalculan sobre una rejilla hexagonal multinivel para el conjunto de zonas semilla, correspondientes a todos los centroides de parcelas residenciales. En la Figura 4.3 se muestra el grado de granularidad de las medidas de accesibilidad, 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 metros[18], para la resolución 9 la arista mide 174 metros.

[16] Discrete Global Grid System.

[17] El 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.

[18] En 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:

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

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

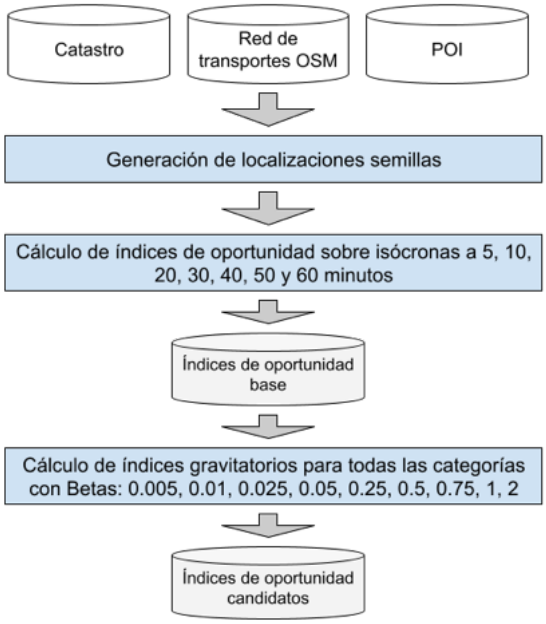
donde para cada elemento o en el universo completo de oportunidades O, con un par de coordenadas (o_x, o_y), definimos una medida de contribución C_k, cuya definición depende de la función de agregación a aplicar en cada familia k (recuento, suma o densidad). $M(X, Y, o_x, o_y, d_{i(m)})$ es una función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

El conjunto de $k = 1, ..., K$ 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.

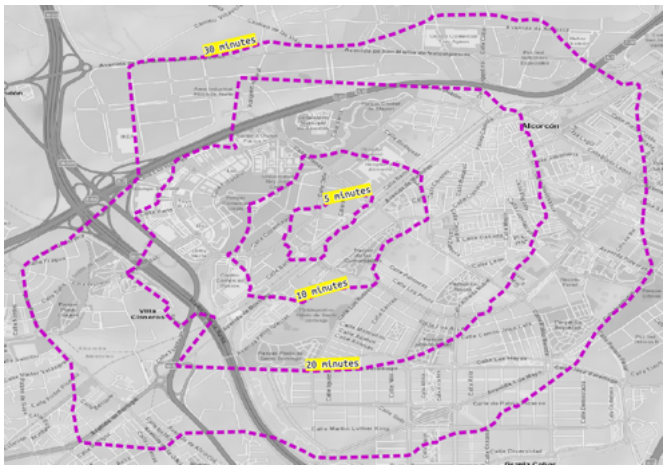
Figura 4.4. Parte I - Generación de índices de accesibilidad candidatos



Fuente: elaboración propia.

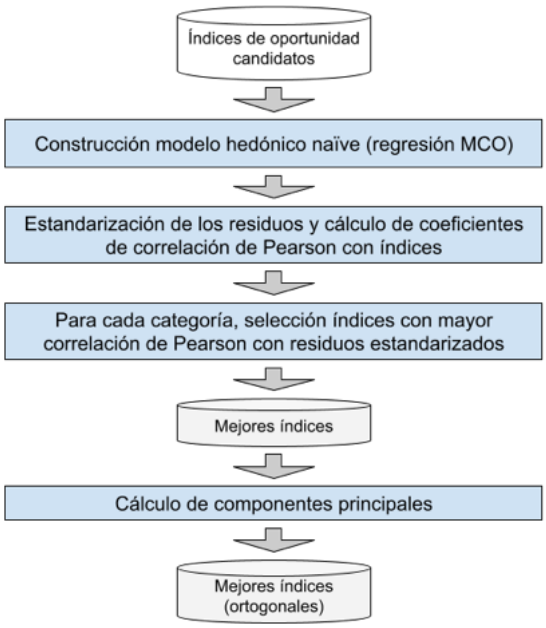
Sobre las ubicaciones de semilla se calculan las áreas isócronas asociadas para la serie de distancias de tiempo en los dos medios de transporte: $d_{i(m=coche)}$ y $d_{i(m=a\ pie)}$. La Figura 4.5 ilustra la forma de los anillos concéntricos definidos alrededor de una semilla específica.

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:

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

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 posible 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 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.

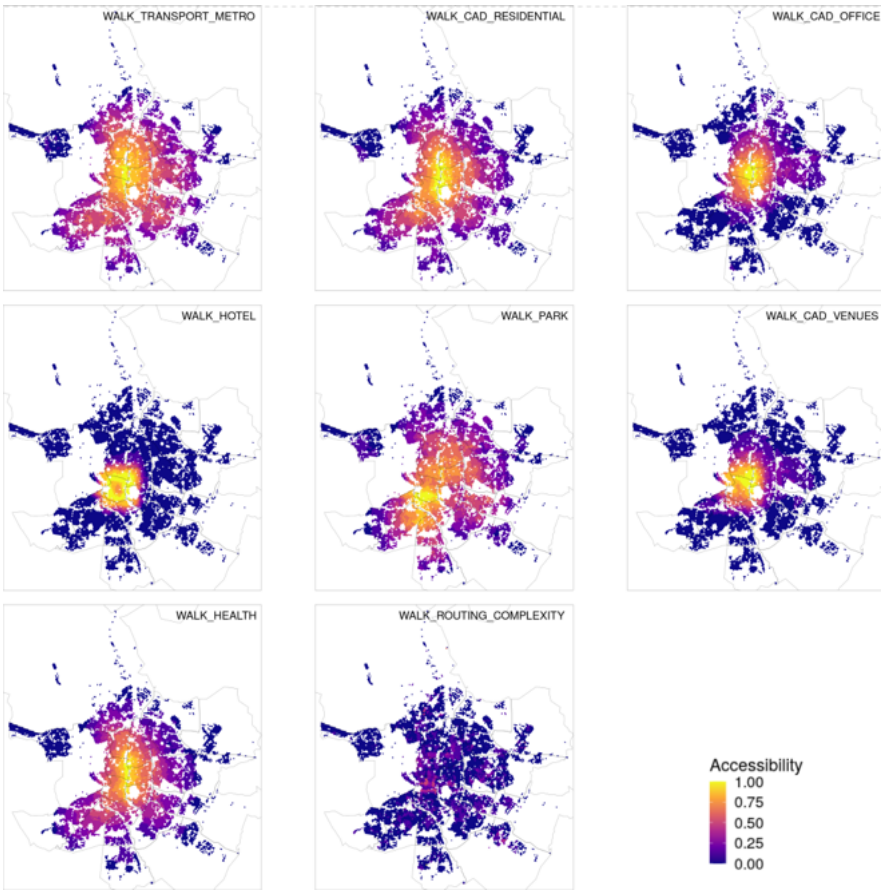
El enfoque heurístico, además, selecciona la mejor configuración de cada índice accesibilidad individual (en este caso las β), con la expectativa de reducir el sesgo espacial del modelo *naïve*. Este procedimiento constituye una versión simplificada de un algoritmo de *boosting* (Friedman y Weinberg, 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 $1 - e^{-\beta_m \cdot 10\ minutos}$).

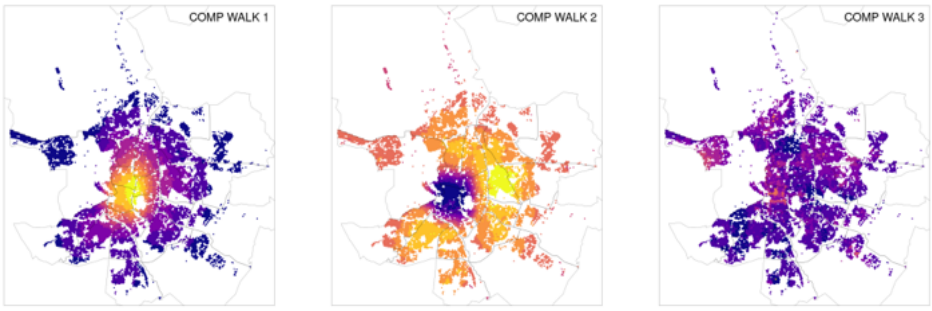
La Figura 4.7 muestra los índices de accesibilidad óptimos A_{nk}^{tm*} para las variables seleccionadas. Los colores más claros indican un mayor acceso a las oportunidades. Por ejemplo, los hoteles (*WALK HOTELS*) están muy concentrados en el centro de la ciudad, mientras que los servicios de salud 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, medido en 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 WALK 1**	**COMP WALK 2**	**COMP WALK 3**	**COMP WALK 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 COMPLEXITY	0.11	0.12	**0.98**	
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 components principales - modo a pie

	Autovalores iniciales			Suma de las cargas al cuadrado			Suma de las cargas al cuadrado rotadas		
Componente	**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 CAR 1**	**COMP CAR 2**	**COMP CAR 3**	**COMP 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 COMPLEXITY	-0.41	-0.20		**-0.89**

Fuente: elaboración propia

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

	Autovalores iniciales			Suma de las cargas al cuadrado			Suma de las cargas al cuadrado rotadas		
Componente	**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 distrito[19] y datos sociodemográficos[20]. 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 esperados[21] 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.

[19] Se incluye la proporción de viviendas en compra y alquiler y número medio de contactos por anuncio en el distrito.
[20] Se incluye el nivel de educación y densidad de población del distrito.
[21] El 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 regularización[22]. 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 CART[23].
- *Random Forests*: este método [24] 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

[22] Se ha utilizado el paquete de R *glmnet* (Simon *et al.*, 2011).
[23] El árbol de tipo CART (Breiman *et al.*, 1984) es un algoritmo de árboles de decisión y regresion 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).
[24] Se 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. R^2: 0,71089, R^2 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	$MedAE = mediana\|e_i\|$
Error absoluto medio porcentual	$MAPE = \frac{1}{N}\sum_{i=1}^{N} \frac{\|e_i\|}{x_i}$
Error absoluto mediano porcentual	$MedAPE = \ mediana\frac{\|e_i\|}{x_i}$
R^2	$R^2 = 1 - \frac{\sigma^2_{error}}{\sigma^2}$

Fuente: elaboración propia

4.3 Resultados

A continuación se evalua 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

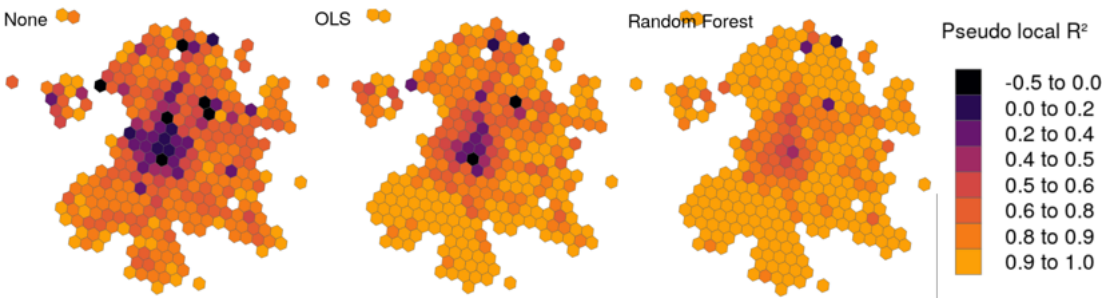
Algoritmo	Accesibilidad		Ninguno
	Val. Cruzada	Val. Cruzada Espacial	Val. Cruzada Espacial
MCO	18.8%	24.1%	43.3%
LERG	18.8%	24.0%	43.3%
Árbol RP	21.9%	26.9%	38.6%
R Forests	10.7%	24.5%	38.5%

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 H3[25]. 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:

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

Figura 4.9. Pseudo R² 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

[25] Se 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:

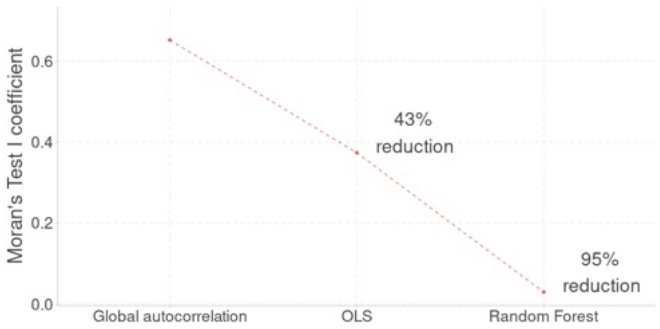
$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad [4.6]$$

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 OSM[26] 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.

[26] POI 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

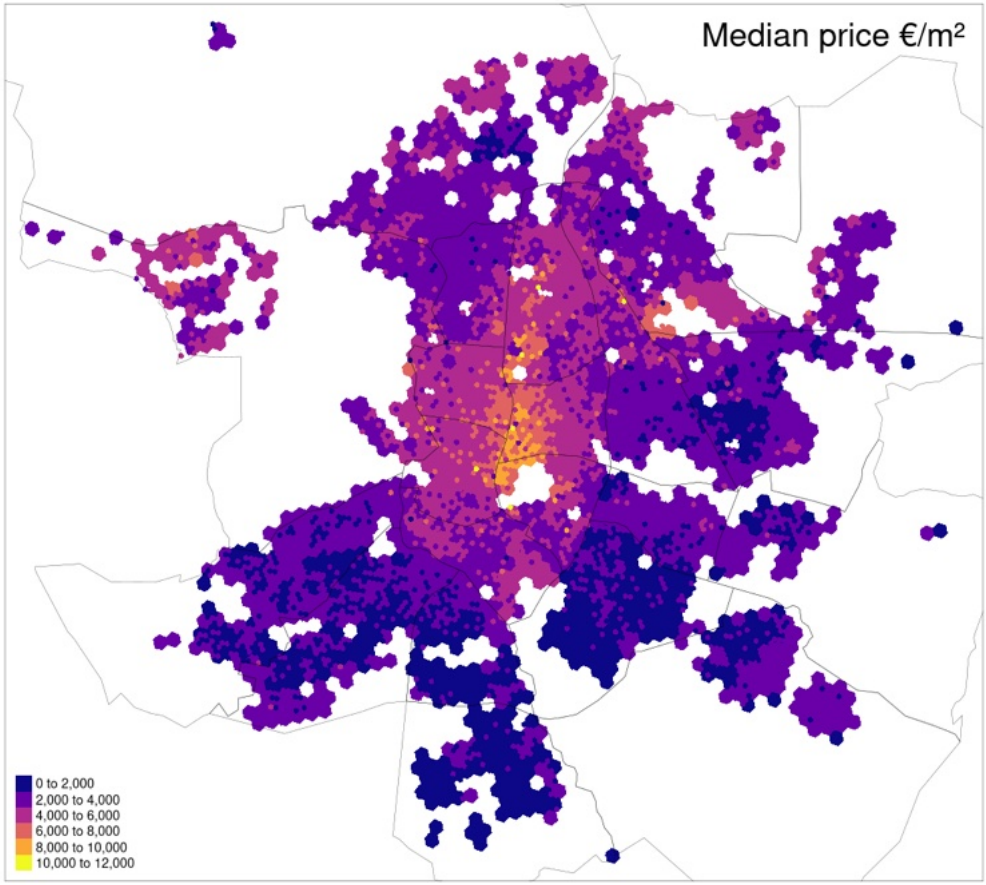
Algoritmos	**Paquete R**	**Parametros**	**Función de pérdida**
MCO	stats	No aplica	Mínimos cuadrados
LERG	glmnet	family = "gaussian", alpha = 0.2659	Mínimos cuadrados
Árbol RP	rpart	cp = 0.002187, minsplit = 8	RMSE
R Forests	ranger	num.trees=96, mtry=37	RMSE

Fuente: elaboración propia

Anexo 4d. Distribución espacial de precios

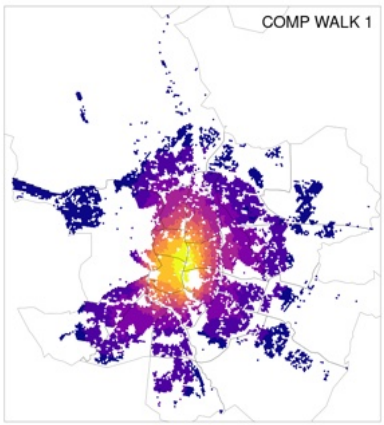
La Figura 4.11 muestra la distribución espacial de los precios por metro cuadrado construido para la ciudad de Madrid, se observa como la zona central ofrece niveles de precios más altos.

Figura 4.11. Precio mediano en €/m²

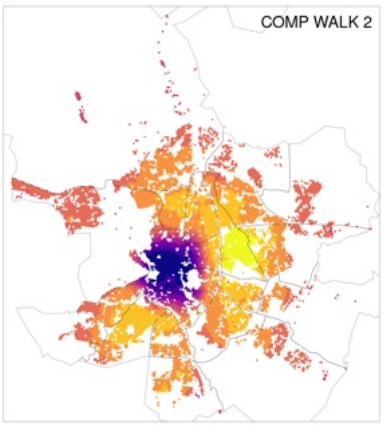


Las 4 gráficas de la Figura 4.12 muestran los patrones espaciales de valores para cuatro componentes de accesibilidad: 3 correspondientes al medio de transporte a pie y 1 para coche.

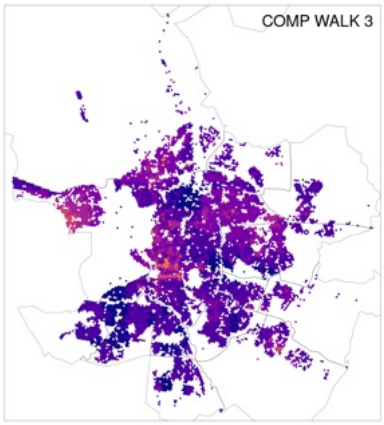
Figura 4.12. Distribución espacial componentes principales de accesibilidad



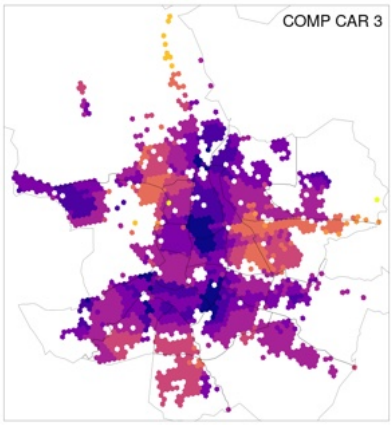
(a) Componente PCA 1 - a pie



(c) Componente PCA 2 - a pie



(b) Componente PCA 3 - a pie



(d) Componente PCA 3 - coche

Anexo 4e. Test de Moran I

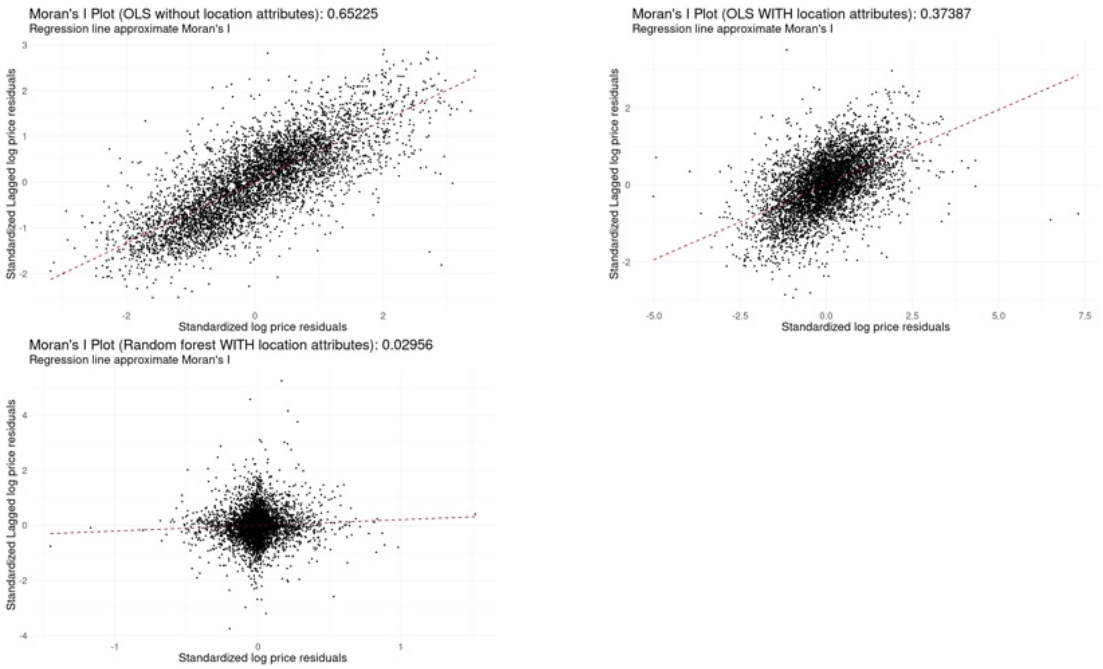
La Tabla 4.14 muestra la autocorrelación espacial sobre los residuos del modelo, calculada sobre el precio mediano por superficie construida para todas las áreas H3 de resolución 8 del municipio de Madrid. La autocorrelación global se estima sobre un modelo que no tiene en cuenta la posición geográfica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

Tabla 4.14. Índice de Moran I para autocorrelación espcial

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



Capítulo 5

Modelo hedónico de oferta

"Las cosas valen lo que se pagan por ellas en una venta."

— Edward Coke, jurista británico

5.1 Introducción

En los capítulos anteriores se ha trabajado en un modelo que represente la relación que guardan los precios y la estructura de pesos muestrales para los colectivos del alquiler y los de oferta (capítulos 2 y 3). La limitación de estos modelos procede de que varias de las fuentes de información sobre las que se trabaja, como el censo o la EPF, no cuentan con el suficiente desglose funcional, geográfico y temporal.

No obstante, dado que ya se conoce la relación funcional general entre poblaciones y precios, y se dispone de registros de oferta desagregados, se pueden mitigar las limitaciones anteriores mediante un modelo hedónico de oferta altamente detallado. Este modelo permitirá trasladar este nivel de desglose a los precios de mercado, calculados según la metodología de los capítulos anteriores.

El modelo de oferta a construir persigue un alto de nivel de fiabilidad, capacidad de ajuste y desglose, para todos los distintos estratos que componen la población y supone una verdadera innovación en el área. Por tanto, los modelos de aprendizaje estadístico son los candidatos ideales a aplicar en la metodología, en particular los modelos de árboles ensamblados de tipo Random Forests, a tenor de los resultados de distintos estudios que demuestran su superioridad para este caso de uso, frente otras técnicas de modelización (Alfaro Navarro *et al.*, 2020; Antipov y Pokryshevskaya, 2012; Graczyk *et al.*, 2010; Truong *et al.*, 2020).

Los modelos de valoración basados en datos de oferta de portales inmobiliarios se han popularizado en las últimas dos décadas, con una numerosa literatura

de casos en distintas geografías. Por ejemplo, en España (Alfaro Navarro *et al.*, 2020; Baldominos *et al.*, 2018; Del Cacho, 2010; Larraz y Poblacion, 2013); en Estambul (Turquía) (Özsoy y Şahin, 2009); en Montreal (Canadá) (Pow *et al.*, 2014); la República Checa (Larraz y Poblacion, 2013); en China (Truong *et al.*, 2020); (Clark y Lomax, 2018) en Reino Unido y (Pérez-Rave *et al.*, 2019) en Colombia.

Una de las primeras referencias en aplicar *Random Forests* en la valoración de vivienda es la de Antipov (Antipov y Pokryshevskaya, 2012), que argumenta su idoneidad comparado con otros métodos. En su estudio trabaja sobre un conjunto de datos de San Petersburgo (Rusia) y compara su rendimiento con 10 algoritmos de aprendizaje automático, evaluando su comportamiento general y en distintos segmentos de la muestra, adicionalmente, propone otras mejoras basadas en la aplicación de coeficientes de corrección en los distintos segmentos de la población. Čeh *et al.* (2018) comparan el desempeño de un modelo hedónico basado en regresión múltiple para viviendas en Lubjiana (Eslovenia), indicando que esta aproximación logra mejores resultados en todas las métricas habituales de evaluación de ajuste y error. Es interesante la construcción de una serie de componentes de accesibilidad, a través de PCA (Pearson, 1901), como variables auxiliares. Por su parte, Hong (2020) aplica *Random Forests* sobre viviendas del barrio de Gangnam en Seúl (Corea del Sur). En su estudio, el modelo de árboles reduce a una cuarta parte el error absoluto del modelo hedónico de regresión. En Noruega, Hjort (2022) estudian las mejoras de precisión al usar modelos de ensamblados de árboles, en este caso comparando el error en la valoración de carteras masivas de inmuebles (AVM).

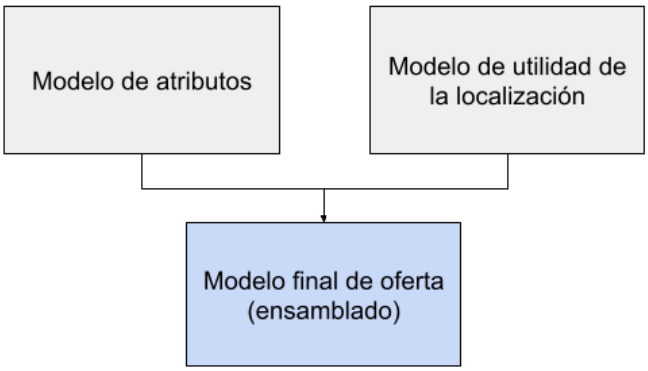
Para el caso español, se han publicado recientemente tres estudios donde se han usado modelos de esta naturaleza. Rico y Taltavull (2021) construyen un modelo de bosques aleatorios sobre un inmuebles de la provincia de Alicante, centrado en aspecto de la explicabilidad; Baldominos (2018) utiliza diferentes técnicas de aprendizaje automático, entra ellas Random Forests, para construir un modelo de regresión de precios de la vivienda para Madrid; Alfaro et al. (2020) construyen un modelo de valoración para 433 municipios, probando distintos modelos ensamblados, y logrando el mejor desempeño de todos con Random Forests.

Sin embargo, estos modelos requieren un establecimiento adecuado de los hiperparámetros del modelo (Antipov y Pokryshevskaya, 2012). Además, debe tenerse en cuenta que la precisión del modelo es sensible a los parámetros establecidos por el analista, según sugieren Prinzie y Van den Poel (2008), que centran su investigación sobre modelos de árboles. Otra cuestión relevante

sobre los árboles de regresión es su todavía limitada adopción en el campo econométrico, principalmente por su menor interpretabilidad que los métodos tradicionales. Si bien, en los últimos años ha habido avances importantes en este sentido (Lundberg *et al.*, 2018; Lundberg y Lee, 2017), con la introducción de técnicas que permiten explicar el comportamiento de los modelos de aprendizaje estadístico.

Como se presentaba en los capítulos 3 y 5, los inductores de precio de la vivienda son múltiples. Hill (2013) los reduce a los dos más relevantes: características y la localización. La metodología seguida para el modelo de oferta propone construir un modelo para cada inductor, junto con un tercer modelo ensamblado que los une, de forma que, se puedan incorporar tanto las contribuciones de los dos aspectos como las de sus interacciones. La Figura 5.1 representa gráficamente el proceso.

Figura 5.1. Diagrama general del modelo hedónico de oferta



Fuente: elaboración propia.

El capítulo se estructura en dos partes: primeramente, se describe el planteamiento metodológico para construir el modelo de oferta, centrado en cual es la mejor forma de elaborarlo, si sobre un modelo único o combinando varios; en segundo lugar, se presentan los resultados obtenidos, centrando la discusión en el aporte de la combinación de modelos, el ajuste espacial, el control de sesgos y la interpretabilidad.

5.2 Metodología

El proceso de creación de modelos hedónicos de precios de la vivienda presenta varias dificultades: la inexistencia de una forma funcional canónica, el comportamiento de mercado de los distintos segmentos de producto (unifamiliar, plurifamiliar) y la influencia de la dimensión espacio-temporal. Para intentar incorporar, en la mayor medida posible, el efecto de estos aspectos, se han construido diferentes modelos hedónicos por cada tipo de inmueble que se "ensamblan" para obtener la mejor estimación.

El concepto de ensamblado, o consenso de modelos, se fundamenta sobre la capacidad de construir un estimador fuerte mediante la agregación de una serie de modelos débiles o base. Estos últimos se pueden construir mediante diversas técnicas (árboles de decisión, redes neuronales, regresiones simples, etc.), para el modelo denominado de ensamblado combine todos los resultados que aproveche los distintos aspectos positivos de los modelos individuales, y compense sus errores (Zhou, 2021). La idea inicial de ensamblado es bastante antigua, y una de sus primeras menciones se puede encontrar en la investigación de Hansen y Salamon (1990), que demostró que la predicción combinada de una serie de modelos mejoraba los resultados de un clasificador individual.

En el campo del aprendizaje estadístico, los ensamblados comenzaron a popularizarse a partir de la década de los 2000, y se han aplicado eficazmente a varios campos como el sector farmacéutico, la banca, los sistemas de recomendación de contenidos o el control del fraude (Seni y Elder, 2010).

La investigación de Schapire (1990) prueba que la precisión de los modelos débiles puede multiplicarse a través de modelos agregados, denominados fuertes. Este trabajo da lugar a la familia de modelos denominados de *boosting* adaptativos, como por ejemplo Adaboost (Freund *et al.*, 1999), que ha sido uno de la más influyentes dentro de esta categoría.

Aunque existen trabajos teóricos sobre los métodos más comunes: *stacking*, *boosting* y *bagging*, no existe un entendimiento completo de los mecanismos subyacentes de los métodos, aunque los resultados empíricos indican que estas aproximaciones no adolecen de problemas de sobreajuste (Zhou, 2021) después de un número suficientemente grande de iteraciones, y en ocasiones, son capaces de reducir el error. Al ser métodos no paramétricos, el rendimiento de estos modelos se realiza mediante el estudio de la descomposición de sesgo-varianza. Para los algoritmos basados en *bagging*, se conoce que son eficaces en la reducción de la varianza, por tanto, ideales para aplicarlo en conjuntos con una alta varianza (este es uno de los motivos de usar Random Forests de forma

extensiva en la presente investigación). Adicionalmente, los modelos de *boosting* son capaces de reducir ambos factores de una manera eficaz (Hastie *et al.*, 2017).

En nuestro caso, se crea un modelo de ensamblado a medida con el objeto de controlar adecuadamente los fenómenos de la varianza y el sesgo (Zhou, 2021), bajo un principio de simplicidad, puesto que, se ha observado que una mayor complejidad en el ensamblado está asociada a resultados más pobres (Graczyk *et al.*, 2010). El método combina dos modelos para reducir de forma secuencial la varianza, controlando el sesgo introducido en cada paso.

Se unen dos modelos anuales, uno basado en atributos y otro basado en la localización, que hacen un total de 27 modelos para la serie histórica, 3 por cada uno de los 9 años de la serie[1]. Al especializar los dos procesos, se intenta evitar que un modelo único pase por alto alguna de las variables relevantes de la muestra, maximizando el aprovechamiento de las contribuciones marginales de las características en los dos ámbitos más importantes: los atributos constructivos y el área en la que se encuentra la vivienda. Desde un punto de vista inmobiliario-urbanístico, esta división se refiere al desglose de los dos atributos fundamentales que forman el precio de la vivienda, el precio del suelo y el precio de la construcción (o vuelo).

Dada la complejidad del conjunto de datos, en términos de atributos y variedad de submercados de la vivienda, se decide utilizar la técnica de modelado de regresión basado en árboles. Se ha optado por el algoritmo *Random Forests*[2] (Breiman, 2001), por su capacidad de gestionar no linealidades, precisión, velocidad de convergencia y menor tendencia al sobreajuste.

Una cuestión importante a tener en cuenta en la evaluación de resultados, es que los modelos de *bagging* no producen modelos sesgados, en el sentido estricto de que la media de los errores tiende a ser cero, pero si comportan otros tipos de sesgos. A este respecto, Breiman (1996) afirma que *Random Forests* reduce la varianza pero no actúa de forma eficaz sobre el sesgo del modelo, puesto que, tiende a modelar correctamente los valores medios (representados por las hojas finales del árbol), pero tiene dificultad para predecir los casos extremos. Se profundizará en esta cuestión en el epígrafe 6.3.2 del próximo capítulo, dedicado a los sesgos en el modelo hedónico final.

[1] Los tres modelos se corresponden al de atributos, localización y ensamblado.

[2] Para este caso se ha utilizado la versión denominada ranger (Wright y Ziegler, 2015), por su capacidad de trabajar eficientemente con grandes volúmenes de datos con alta dimensionalidad.

5.2.1 Modelo hedónico de características

El modelo hedónico de características construye la relación entre el precio y los atributos de la vivienda. Como se muestra en la Tabla 5.1 y en la Tabla 5.2, los 43 atributos atienden a características físicas del inmueble y de la finca, dinámicas de mercado de la zona, calidad y estado de conservación del inmueble.

Tabla 5.1. Variables modelo de atributos

Categoría	Variable	Fuente	Descripción	Inmueble
Area	CLUSTER	calculado	Tipo de zona siguiendo una clasificación propia	Unifamiliar
Calidad	CADASTRALQUALITYID	catastro	Calidad de la construcción	Ambos
Edificio	AMENITYID	idealista	Tipo de instalaciones de la finca	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
	CONSTRUCTIONYEAR	catastro	Año de construcción de la propiedad	Ambos
	DWELLING_COUNT	catastro	Número de inmuebles en la finca	Plurifamiliar
	FLOOR_POSITION	idealista	Posición del piso dentro del edificio	Plurifamiliar
	HASDOORMAN	idealista	Tiene portero	Unifamiliar
	HASGARDEN	idealista	Tiene jardín	Unifamiliar
	HASLIFT	idealista	Tiene ascensor	Plurifamiliar
	HASSWIMMINGPOOL	idealista	¿Tiene su edificio una piscina?	Ambos
	MAXBUILDINGFLOOR	idealista	Número de pisos en edificio	Ambos
Fecha	PERIOD	idealista	Código mes año en formato YYYYMM	Ambos
Mercado	CHANNELID	idealista	Canal de comercialización (agencia / particular)	Ambos
	LEADS_RESIDENTIAL	idealista	Número de contactos medios en la zona	Unifamiliar
	ONMARKET_RENT	idealista	Número de inmuebles en alquiler en la zona	Unifamiliar
	ONMARKET_SALE	idealista	Número de inmuebles en venta en la zona	Unifamiliar
	RENTSALE_RATIO	idealista	Proporción de inmuebles en alquiler / venta	Unifamiliar

Fuente: elaboración propia

Para evitar, en buena medida, el sesgo ocasionado por no incluir la información de la zona (precio del suelo), se incluye la variable denominada *Cluster*, que indica la categoría de zona a la que pertenece la seccion censal en la que se encuentra el

anuncio. Estas categorías se estiman de forma automática mediante un algoritmo de análisis *cluster*, descrito en el Anexo 5a de este capítulo.

Otra aportación original de este modelo de atributos es la incorporación de las características de competencia del mercado. Existen evidencias empíricas, como en el modelo propuesto por Fuss y Koller (2016), que señalan la capacidad predictiva de las variables que describen las dinámicas de los mercados locales.

Tabla 5.2. Variables modelo de atributos (continuación)

Categoría	Variable	Fuente	Descripción	Inmueble
Estructura	BATHNUMBER	idealista	Número de baños	Unifamiliar
	BEDROOMNUMBER	idealista	Número de dormitorios	Unifamiliar
	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Unifamiliar
	ENERGYCERTIFICATIONID	idealista	Código de certificado energético	Unifamiliar
	FLATLOCATION	idealista	Indica si el piso es interior o exterior	Plurifamiliar
	FLOOR	idealista	Planta en la que está el inmueble	Plurifamiliar
	GARAGETYPEID	idealista	Tipo de garaje	Unifamiliar
	HASAIRCONDITIONING	idealista	¿Tiene aire acondicionado?	Ambos
	HASANNEX	idealista	El piso tiene anejos	Plurifamiliar
	HASBALCONY	idealista	Tiene balcón	Plurifamiliar
	HASBOXROOM	idealista	Tiene trastero	Unifamiliar
	HASEASTORIENTATION	idealista	Está orientado al este	Unifamiliar
	HASNORTHORIENTATION	idealista	Está orientado al norte	Unifamiliar
	HASPARKINGSPACE	idealista	Tamaño del garaje	Unifamiliar
	HASSOUTHORIENTATION	idealista	Está orientado al sur	Unifamiliar
	HASTERRACE	idealista	Tiene terrazas	Ambos
	HASWARDROBE	idealista	Tiene armarios empotrados	Ambos
	HASWESTORIENTATION	idealista	Está orientado al oeste	Unifamiliar
	ISDUPLEX	idealista	Es un dúplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
	PLOTOFLAND	idealista	Tamaño de la parcela unifamiliar	Unifamiliar
	ROOMNUMBER	idealista	Número de habitaciones	Ambos
	USABLEAREA	idealista	Área útil	Unifamiliar

Fuente: elaboración propia

Debido a que es un modelo de regresión, cuya variable objetivo tiene una gran variabilidad, dispersión y no-normalidad, se opta la modalidad de árboles de

regresión cuantílicos. Adicionalmente, se usan los pesos poblacionales para que el modelo resultante tenga en cuenta la distribución real de la población. Es posible configurar distintos hiperparámetros para mejorar el ajuste de los modelos, tales como:

- *num.trees*: número de árboles utilizados.
- *min.node.size*: tamaño mínimo de nodo final del árbol, en nuestro caso se usa 8 para evitar el sobreajuste.
- *quantreg*: si está activo, el modelo realiza una regresión cuantílica sobre el método *Random Forest* (Meinshausen, 2006), que es una generalización del método original propuesto por Breiman. La ventaja esta clase de regresión es que proporciona información sobre la distribución condicional completa de la variable de respuesta, y no solo sobre la media condicional como en el método de regresión habitual. Su principal ventaha es que este método proporciona una aproximación, no paramétrica y precisa, de regresión cuantílica para problemas con un gran número de variables.
- *importance*: modo de cálculo de la importancia, y representa la forma en la que se mide la capacidad de reducción de la entropía de cada variable. Para este caso, se establece el parámetro *impurity* refierido a la reducción de la varianza.
- *mtry*: número de variables sobre las que se evalúa hacer la división en cada nodo. Por defecto es la raíz cuadrada del número de variables, y nunca debe superar el número total de ellas.
- *maximal.tree.depth*: Indica la profundidad máxima del árbol a construir. En este caso no se restringe este valor.

Dado que existen múltiples combinaciones posibles, se fija *mtry* a la raíz cuadrada del número de parámetros, y se realiza una evaluación de distintas combinaciones del resto de parámetros mediante la técnica de *grid search*[3]. Adicionalmente, se ha aplicado un remuestreo de tipo validación cruzada (LeCun *et al.*, 2015), que ha demostrado ser un criterio consistente para este fin (Yang, 2007).

El criterio de selección de la configuración es aquel que haga máximas las métricas clave (Kuhn *et al.*, 2018), en este caso: el coeficiente de determinación R^2 y el error del modelo, medido como el error cuadrático medio[4] (RMSE). Debido a las diferencias entre los inmuebles de tipo unifamilior y plurifamiliar, se han estimado modelos diferentes por tipología.

Las configuraciones probadas para el modelo ensamblado corresponden al año

[3]La técnica de *Grid Search* consiste en la prueba sistemática de múltiples conmbinaciones de parámetros para un algoritmo de aprendizaje automático. Su objetivo es encontrar la configuración que mejor funciona, en base a unas métricas de rendimiento del modelo.

[4]Tanto el R^2 como el error se miden sobre el conjunto *out of bag*, que ya se analizó en el epígrafe con el mismo nombre del Anexo II, en el capítulo 3.

2019 (se ha tomado este año por ser el más informado de la serie), y se detallan en la Tabla 5.3. Se observa que la precisión aumenta a medida que aumenta aumenta el número árboles y se reduce el tamaño mínimo de nodo. Se decide utilizar 200 árboles con un tamaño mínimo de nodo de 8 instancias. Para unifamiliar, se observa 150 árboles ofrecen un nivel de ajuste similar a 200, sin degradación apreciable en RMSE.

Tabla 5.3. Resultados de la búsqueda de hiperparametros

		Plurifamiliar			Unifamiliar		
num.trees	min.node.size	RMSE	R2	Tiempo	RMSE	R2	Tiempo
200	8	10.34	76.6%	13.84	3.37	81.5%	23.28
150	8	10.44	76.4%	10.39	3.38	81.5%	19.16
100	8	10.60	76.0%	9.68	3.44	81.1%	15.52
50	8	10.95	75.2%	3.69	3.66	79.9%	11.58
200	16	11.55	73.9%	12.59	3.66	79.9%	22.26
150	16	11.63	73.7%	9.88	3.71	79.7%	18.31
100	16	11.77	73.4%	7.83	3.78	79.3%	14.94
50	16	12.05	72.7%	3.58	3.86	78.9%	11.12
10	8	12.94	70.7%	1.60	4.57	75.0%	9.05
200	32	13.19	70.1%	11.85	4.22	76.9%	21.15
150	32	13.25	70.0%	9.46	4.29	76.5%	17.93
100	32	13.41	69.6%	5.91	4.38	76.0%	14.47
50	32	13.66	69.1%	3.48	4.50	75.3%	11.32
10	16	13.85	68.6%	1.54	4.77	73.9%	8.43
10	32	15.18	65.6%	1.51	5.26	71.1%	8.14

Fuente: elaboración propia

Los hiperparámetros seleccionados varían ligeramente según el tipo de inmueble, como muestra la Tabla 5.4. Las diferencias residen en el número de árboles utilizado y el número de variables para las divisiones (*mtry*), en unifamiliar ambos parámetros son ligeramente mayores por su diversidad.

Tabla 5.4. Hiperparámetros modelo de atributos

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	4	TRUE
Unifamiliar	200	impurity	8	5	TRUE

Fuente: elaboración propia

El error se pondera con los elevadores muestrales para priorizar la reducción del error de las observaciones con mayor representación. Por otra parte, aunque no se ha restringido la profundidad máxima del árbol, se ha utilizado un alto número de árboles para limitar el potencial sobreajuste debido al factor anterior.

5.2.2 Modelo hedónico de utilidad de la localización

El uso de información espacial aporta una notable mejora en la calidad del modelado hedónico de índices del precio de la vivienda, y aunque que no hay un método canónico para aplicarla, el aspecto clave es utilizar unidades de análisis suficientemente pequeñas (Hill y Scholz, 2018).

De forma estricta, este modelo no es un modelo hedónico, sino que es un modelo de corrección del sesgo del modelo de atributos, mediante el uso de variables de localización. Por tanto, la variable objetivo del modelo es la proporción entre el precio real por metro cuadrado y la estimación de precio según el modelo de atributos, es decir:

$$y_i = \frac{p_i}{\hat{p_i}^{atributos}} \quad [5.1]$$

donde y_i representa la variable objetivo para la observación i, p_i el precio real y $\hat{p_i}^{atributos}$ el precio estimado por el modelo de atributos.

Como en el resto de casos, se ha utilizado *Random Forests* con los hiperparámetros recogidos en la Tabla 5.5. En este caso, el valor de *mtry* en plurifamiliares es mayor en el modelo de características, por tener un mayor número de covariables.

Tabla 5.5. Hiperparámetros modelo de localización

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	5	TRUE
Unifamiliar	200	impurity	8	5	TRUE

Fuente: elaboración propia

Se han utilizado un conjunto de 18 variables relacionadas con la utilidad marginal asociada a la zona (utilidad/accesibilidad), que se obtienen a través del método presentado en el capítulo 4. Estas variables resumen qué instalaciones o servicios dispone una zona, teniendo en cuenta que los individuos que habitan en cada vivienda se desplazan tanto a pie como en transporte privado.

En la Tabla 5.6, se describen las 32 variables utilizadas, clasificadas en diferentes categorías: accesibilidad, datos básicos del edificio, subtipología y atributos sociodemográficos.

Tabla 5.6. Variables modelo de localización

Categoría	Variable	Fuente	Descripción	Inmueble
Accesibilidad	COMP CAR 1 .. 9	calculado	Accesibilidad en coche	Ambos
	COMP WALK 1 .. 9	calculado	Accesibilidad caminando	Ambos
Area	CLUSTER	calculado	Tipo de zona siguiendo una clasificación propia	Unifamiliar
Edificio	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
Estructura	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Ambos
	ISDUPLEX	idealista	Es un duplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
Mercado	LEADS RESIDENTIAL	idealista	Número de contactos medios en la zona idealista	Plurifamiliar
	ONMARKET RENT	idealista	Número de inmuebles en alquiler en la zona	Plurifamiliar
	ONMARKET SALE	idealista	Número de inmuebles en venta en la zona	Plurifamiliar
	RENTSALE RATIO	idealista	Proporción de número de inmuebles en alquiler versus en compra (en oferta)	Plurifamiliar
Zona	AGE 3	INE	Porcentaje de mayores en la sección censal	Ambos
	DENSITY	INE	Densidad de población del tramo censal	Ambos
	EDUCATION 3	INE	Porcentaje de personas con estudios superiores en la sección censal	Ambos
	POPULATION MUNICIPALITY	INE	Población del municipio	Ambos
	RATE FOREIGN	INE	Tasa de extranjeros en la sección censal	Ambos

Fuente: elaboración propia

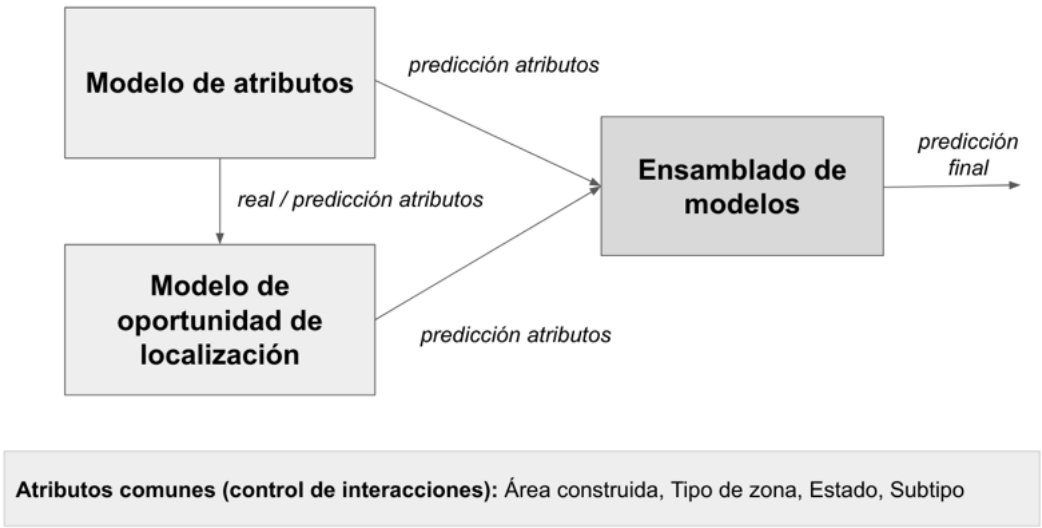
Los atributos zonales se complementan con datos socio-demográficos de la zona y con la tipología de zona calculada (*CLUSTER*). Esto permite explicar los aspectos que no cubren los indicadores de accesibilidad, por ejemplo, dos áreas del norte y sur de Madrid pueden tener precios de suelo diferentes aún teniendo una misma configuración en términos de accesibilidad.

5.2.3 Modelo ensamblado contra modelo único

Se decide construir un modelo final sobre un ensamblado de modelos de características y zona en base a los estudios que demuestran empíricamente que este enfoque produce mejores resultados que los modelos individuales (Graczyk *et al.*, 2009; Hashem, 1997; Krogh y Vedelsby, 1994; Opitz y Shavlik, 1996). En todo caso, es necesario tener en cuenta que, en ocasiones, pueden producirse fenómenos de sesgo (Graczyk *et al.*, 2010).

La forma funcional del ensamblado es semilogarítmica y combina, de forma aditiva, las predicciones del modelos individuales. La idea general, expresada en la Figura 5.2, se basa en que los dos primeros modelos se centran en la reducción de la varianza, cada uno usando un aspecto específico de la información disponible, y el modelo final se enfoca en eliminar el sesgo introducido por cada uno de los anteriores.

Figura 5.2. Esquema general del ensamblado de modelos



Fuente: elaboración propia.

Dado que es posible encontrarse grados específicos de interacción de las variables de cada modelo, porque la influencia del número de habitaciones puede ser distinta entre unos barrios y otros, se usan un conjunto de variables comunes para controlar estas interacciones. De esta forma, se evita que la generación de sesgo, la heterogeneidad espacial o las variables omitidas relevantes.

Los predictores utilizados, se han seleccionado en base a las conclusiones de distintos estudios que los identifican como los más representativos. Véase, por ejemplo, los resultados de Rico y Taltavull (2021) o de Clark et al. (2018). Las

variables utilizadas han sido:

- Tipo de zona: a través de las variables *Cluster* y zona idealista, se utiliza en el modelo final para corregir los sesgos del modelo de localización.
- Superficie útil.
- Estado de la construcción: nueva, segunda mano reformada y segunda mano sin reformar.
- Subtipo de vivienda unifamiliar: solo para esta tipología, necesaria para diferenciar el comportamiento del precio entre adosados, pareados y viviendas independientes.

El uso de atributos zonales en este modelo permite corregir los sesgos espaciales del modelo de localización debidos a variables omitidas. De tal forma que se puede entender que el modelo de localización captura las interacciones generales sobre las áreas cercanas del inmueble, y el modelo ensamblado final, corrige cuestiones desviaciones entre el precio del suelo estimado usando la localización real (se asume como una corrección de sesgo basada en variables ficticias de zona).

Para el último paso, se usan 7 variables, descritas en detalle en la Tabla 5.7, más del subtipo en el caso de vivienda unifamiliar, que permite mejorar la precisión del modelo ensamblado para este caso.

Tabla 5.7. Variables modelo ensamblado

Categoría	Variable	Fuente	Descripción	Inmueble
Modelo	MODEL 1 PREDICTIONS	interno	Predicción modelo de atributos	Ambos
	MODEL 2 PREDICTIONS	interno	Predicción modelo de localización	Ambos
Area	CLUSTER	interno	Tipo de zona - clasificación interna	Ambos
	LOCATIONID	idealista	Código de área idealista	Ambos
Estructura	CONSTRUCTEDAREA	idealista	Área construida	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de unifamiliar	Unifamiliar

Fuente: elaboración propia

A continuación se describe el modelo ensamblado de forma funcional. Para facilitar su lectura se expresa modelo lineal. La siguiente expresión muestra el modelo de atributos, cuya variable dependiente es el logaritmo del precio por metro cuadrado en euros $log(Precio\ m^2)$:

$$log(P_{atributos}) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot atributo_i + \sum_{k=1}^{n} \beta_k \cdot comunes_i + \varepsilon_a \quad [5.2]$$

donde $atributo_i$ se refiere las covariables de tipo atributo para la observación i, y $comunes_i$ los predictores comunes.

En el modelo de utilidad, las variables independientes principales son los atributos comunes y los atributos del modelo de accesibilidad para cada zona. En este caso, la variable objetivo representa la relación entre el precio real y el precio predicho por el modelo de atributos:

$$R_{utilidad} = \frac{p}{\hat{P}_{atributos}} = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot utilidad_i + \sum_{k=1}^{n} \beta_k \cdot comunes_i + \varepsilon_o \qquad [5.3]$$

donde $utilidad_i$ se refiere las covariables específicas para el modelo de utilidad de la localización.

Finalmente, el modelo ensamblado combina linealmente la contribución de los dos modelos individuales según la expresión:

$$log(P) = \beta_0 + \beta_1 \cdot log(\hat{P}_{atributos}) + \beta_1 \cdot log(\hat{R}_{utilidad}) + \varepsilon_e \qquad [5.4]$$

Los hiperparámetros aplicados al algoritmo *Random Forests*, como muestra la Tabla 5.8, se mantiene casi toda la configuración excepto el parámetro *mtry* debido al reducido número de atributos.

Tabla 5.8. Hiperparámetros modelo ensamblado

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	2	TRUE
Unifamiliar	200	impurity	8	2	TRUE

Fuente: elaboración propia

Para validar la eficacia del modelo ensamblado, se ha desarrollado un modelo que denominaremos único que se tomará como referencia para evaluar el aporte del enfoque ensamblado. El algoritmo utilizado es también *Random Forests*, sobre el que se desarrollan dos modelos diferentes, uno por tipología de vivienda. En ambos casos, la variable a predecir es el logaritmo del precio por metro cuadrado útil, y de forma funcional podría expresarse como:

$$log(P) = \beta_0 + \sum_{i=1}^{n} \beta_i atributos_i + \sum_{i=1}^{n'} \beta'_i \cdot utilidad_i + \sum_{i=1}^{n''} \beta''_i \cdot mercado_i + \varepsilon_e \qquad [5.5]$$

El modelo único cuenta con un mayor número de variables (65), al incorporar información de todos los aspectos del inmueble, como las características

estructurales de la vivienda, los datos de localización o las dinámicas del mercado inmobiliario. El detalle completo de variables utilizadas se presenta en la Tabla 5.9 y la Tabla 5.10.

Tabla 5.9. Variables modelo único

Categoría	Variable	Fuente	Descripción	Inmueble
Accesibilidad	COMP CAR 1 .. 9	calculado	Accesibilidad en coche	Ambos
	COMP WALK 1 .. 9	calculado	Accesibilidad caminando	Ambos
Calidad	CADASTRALQUALITYID	catastro	Calidad de la construcción	Ambos
Edificio	AMENITYID	idealista	Tipo de instalaciones de la finca	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
	CONSTRUCTIONYEAR	catastro	Año de construcción de la propiedad	Ambos
	DWELLING COUNT	catastro	Número de inmuebles en la finca	Plurifamiliar
	FLOOR POSITION	idealista	Posición del piso dentro del edificio	Plurifamiliar
	HASDOORMAN	idealista	Tiene portero	Unifamiliar
	HASGARDEN	idealista	Tiene jardín	Unifamiliar
	HASLIFT	idealista	Tiene ascensor	Plurifamiliar
	HASSWIMMINGPOOL	idealista	¿Tiene su edificio una piscina?	Ambos
	MAXBUILDINGFLOOR	idealista	Número de pisos en edificio	Ambos
Fecha	PERIOD	idealista	Fecha en formato YYYYMM	Ambos
Mercado	CHANNELID	idealista	Canal de comercialización	Ambos
	LEADS RESIDENTIAL	idealista	Número de contactos en zona	Ambos
	ONMARKET RENT	idealista	Inmuebles en alquiler en zona	Ambos
	ONMARKET SALE	idealista	Inmuebles en venta	Ambos
	RENTSALE RATIO	idealista	Proporción alquiler/venta	Ambos
Zona	AGE 3	INE	Pct. de mayores	Ambos
	DENSITY	INE	Densidad de población	Ambos
	EDUCATION 3	INE	Pct. personas con estudios superiores	Ambos
	POPULATION MUNICIPALITY	INE	Población del municipio	Ambos
	RATE FOREIGN	INE	Pct. Extranjeros	Ambos

Fuente: elaboración propia

Tabla 5.10. Variables modelo único (continuación)

Categoría	Variable	Fuente	Descripción	Inmueble
Estructura	BATHNUMBER	idealista	Número de baños	Unifamiliar
	BEDROOMNUMBER	idealista	Número de dormitorios	Unifamiliar
	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Ambos
	ENERGYCERTIFICATIONID	idealista	Código de certificado energético	Unifamiliar
	FLATLOCATION	idealista	Indica si el piso es interior o exterior	Plurifamiliar
	FLOOR	idealista	Planta en la que está el inmueble	Plurifamiliar
	GARAGETYPEID	idealista	Tipo de garaje	Unifamiliar
	HASAIRCONDITIONING	idealista	¿Tiene aire acondicionado?	Ambos
	HASANNEX	idealista	Con anejos (garaje o trastero)	Plurifamiliar
	HASBALCONY	idealista	Tiene balcón	Plurifamiliar
	HASBOXROOM	idealista	Tiene almacenamiento / Trastero	Unifamiliar
	HASEASTORIENTATION	idealista	Está orientado al este	Unifamiliar
	HASNORTHORIENTATION	idealista	Está orientado al norte	Unifamiliar
	HASPARKINGSPACE	idealista	Tamaño del garaje	Unifamiliar
	HASSOUTHORIENTATION	idealista	Está orientado al sur	Unifamiliar
	HASTERRACE	idealista	Tiene terrazas	Ambos
	HASWARDROBE	idealista	Tiene armarios empotrados	Ambos
	HASWESTORIENTATION	idealista	Está orientado al oeste	Unifamiliar
	ISDUPLEX	idealista	Es un duplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
	PLOTOFLAND	idealista	Tamaño de la parcela unifamiliar	Unifamiliar
	ROOMNUMBER	idealista	Número de habitaciones	Ambos
	USABLEAREA	idealista	Área útil	Unifamiliar

Fuente: elaboración propia

Los hiperparámetros aplicados al algoritmo *Random Forests* son similares a los usados en los ensamblados, como muestra la Tabla 5.11. Con la excepción del parámetro *mtry* que se establece en 7, debido al mayor número de variables utilizadas.

Tabla 5.11. Hiperparámetros modelo ensamblado

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	7	TRUE
Unifamiliar	150	impurity	8	7	TRUE

Fuente: elaboración propia

5.3 Resultados

La metodología hedónica de oferta tiene como objetivo producir un modelo general y preciso, por tanto, se evaluará la capacidad de generalización y ajuste de los modelos construidos. Adicionalmente, se estudiará si la incorporación del enfoque ensamblado produce los beneficios esperados, junto con el papel que juegan las distintas variables en la explicación del precio.

Sobre los 36 modelos finales construidos[5], se estudia cuál es la aproximación que mejor resultados ofrece utilizar (ensamblado o único), a través de su grado de ajuste, forma de los errores de predicción y distribución geográfica de los mismos.

5.3.1 Medidas para evaluar la calidad de los modelos

Existe un amplio número de métricas orientadas disponibles para evaluar la calidad de un modelo de precios de la vivienda. Steurer *et al.* (2021) identifican 49 medidas, de las que recomiendan 7, que cubren varios aspectos de análisis sobre los residuos: sesgo, desviación absoluta, ratio de desviación absoluta, desviación cuadrática, ratio sobre desviación cuadrática y porcentaje de errores fuera de rango, medidas sobre cuantiles. Sobre esta base, se han tomado las métricas recogidas en la Tabla 5.12, complementadas con otras de ajuste espacial y sesgo (presentadas en el epígrafe 3.4). El número de medidas debe ser abundante para asegurar la robustez de los resultados (Steurer *et al.*, 2021).

Tabla 5.12. Medidas de error de modelos

Medida	Fórmula
Raíz cuadrada del error medio cuadrático medio	$RMSE = \frac{1}{N}\sum_{t=1}^{N}\sqrt{e_i^2}$
Raíz cuadrada del error medio cuadrático medio normalizado	$NRMSE = \frac{1}{\overline{x}}\frac{1}{N}\sum_{i=1}^{n}\sqrt{e_i^2}$
Error medio absoluto	$MAE = \frac{1}{N}\sum_{i=1}^{N}e_i^2$
Error medio absoluto	$MAE = \frac{1}{N}\sum_{i=1}^{N}e_i^2$
Error mediano absoluto	$MedAE = median\|e_i\|$
R^2	$R^2 = 1 - \frac{\sigma_e^2}{\sigma^2}$
Error medio en porcentaje	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\|e_i\|}{x_i}$
Error mediano en porcentaje	$MEDAPE = mediana\frac{\|e_i\|}{x_i}$

Fuente: elaboración propia

e_i representan los errores de estimación y N el tamaño muestral

[5] Los 36 modelos proceden de las combinaciones entre modelo ensamblado y único, los dos tipos de vivienda y 9 años.

El estudio del ajuste de los modelos se ha realizado en términos monetarios, en linea con la investigación de Pérez-Rave *et al.* (2019), que indica que el análisis de errores sobre la variable transformada[6], a menudo, puede ofrecer resultados ficticios desde un punto de vista estadístico.

Adicionalmente, el sesgo de los modelos estimados representa de la desviación entre la esperanza de los valores predichos y los observados[7]. De forma intuitiva, se puede calcular como la media de las desviaciones de las estimaciones del modelo, que en este caso se realiza de varias formas: como la desviación en términos absolutos, en porcentaje y como curtosis de la distribución de precios estimados. En particular, se ha utilizado la siguiente expresión:

$$Bias(\hat{f}(x)) = E[\hat{f}(x)] = \hat{f}_{observado}(x) - \hat{f}_{estimado}(x) \quad [5.6]$$

donde $\hat{f}(x)$ un estimador sobre el conjunto s, y E representa la esperanza de la desviación de dicho estimador, entre datos observados y predichos.

Todas las métricas de ajuste y error utilizadas, se han calculado sobre las muestras denominadas OOB de los modelos *Random Forests*. Para más información véase el Anexo 3b del capítulo 3.

5.3.2 Resultados de ajuste del modelo

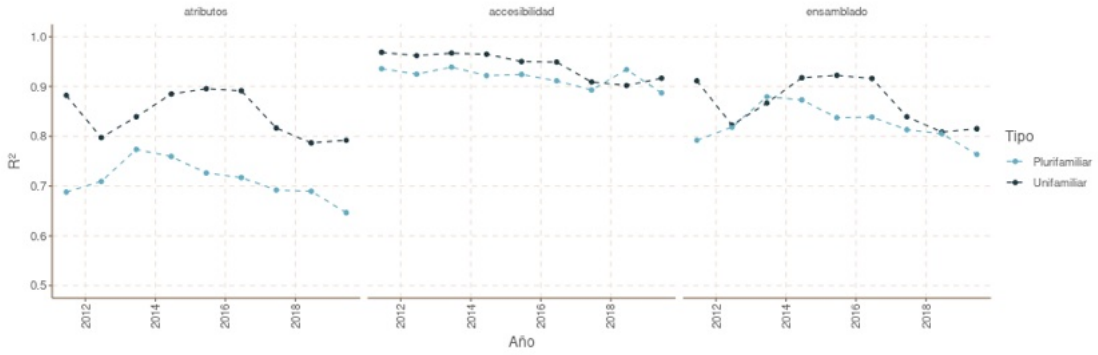
El planteamiento ensamblado pretende que los modelos individuales (atributos y localización) reduzcan la varianza lo máximo posible y, que al combinarlo, se ajusten sus sesgos particulares.

La Figura 5.3 muestra como se comporta el coeficiente de determinación R^2 de los tres procesos, observándose que los modelos individuales capturan buena parte del espectro de la varianza. Con el objetivo de facilitar la correcta interpretación de las medidas de ajuste de los modelos, se recuerda que el coeficiente de determinación del modelo de utilidad se refiere al ajuste sobre razón entre el precio real y el predicho por el modelo de atributos.

[6] A través de una transformación logarítmica o potencial, como puede ser el logaritmo del precio por superficie útil.

[7] El estimador al que se refiere puede ser el precio de la vivienda, los errores del modelo u otra medida de interés relacionada con el modelo.

Figura 5.3. Niveles de ajuste por modelos



Fuente: elaboración propia.

El ajuste medio del modelo de atributos es más bajo debido a que la ubicación geográfica exacta es un determinante importante del precio, en todo caso, el ajuste es satisfactorio al incorporar una mínima especificación zonal, procedente del atributo código de clúster.

La Tabla 5.13 indica que el comportamiento del R^2 es relativamente estable en el tiempo, aunque con cierta degradación en 2012. Se observa que, a partir de 2016, el peor rendimiento del modelo de atributos se corrige eficazmente con el ensamblado. Además, en general, los niveles de ajuste son más altos en unifamiliar que en plurifamiliar, probablemente a consecuencia de una mayor heterogeneidad de este tipo de vivienda.

Tabla 5.13. Ajuste de los modelos en R² desglosado por tipologías

Tipo	Modelo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Unifamiliar	atributos	88.2%	79.7%	83.9%	88.5%	89.6%	89.1%	81.6%	78.7%	79.2%
	accesibilidad	96.9%	96.2%	96.7%	96.5%	95.0%	94.9%	90.9%	90.2%	91.7%
	ensamblado	91.1%	82.2%	86.7%	91.8%	92.2%	91.6%	83.9%	80.8%	81.5%
Plurifamiliar	atributos	68.8%	70.9%	77.4%	75.9%	72.6%	71.7%	69.2%	69.0%	64.6%
	accesibilidad	93.6%	92.5%	93.9%	92.2%	92.4%	91.2%	89.3%	93.4%	88.7%
	ensamblado	79.2%	81.8%	87.9%	87.3%	83.7%	83.9%	81.3%	80.5%	76.4%

Fuente: elaboración propia

Desde un punto de vista comparativo, sobre la medida de R^2, el modelo muestra un rendimiento en un orden similar con otros modelos de precio de alquiler de la literatura, con valores (caso ensamblado) entre 0,76 y 0,92. Aunque es necesario recordar que la escala de esta medida depende mucho de varios factores como la fuente de la calidad de la información o el algoritmo utilizado, se toman como referencia los casos de: Chung (2015) que presenta un 0,98, con un modelo

semilogarítmico; Löch (2010) consigue un 0.856; Fuss y Koller (2016) un 0,883, y Clark y Lomax (2018) logran un 0,69.

En términos de error, la Tabla 5.14 muestra errores medios crecientes en los dos tipos de vivienda, en parte debido a que los precios son crecientes. En el valor normalizado, las viviendas plurifamiliares han un error estable en torno al 17%, mientras que, las unifamiliares ofrecen errores con mucha más fluctuación.

Tabla 5.14. RMSE (absoluto) y NRMSE (normalizado), modelo ensamblado

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Absoluta	Plurifamiliar	28,13	24,59	18,84	20,04	25,30	29,18	34,43	34,54	38,77
	Unifamiliar	15,00	20,87	16,61	12,19	11,61	12,36	17,78	21,70	22,04
Normalizada	Plurifamiliar	17.0%	15.7%	12.8%	13.5%	16.3%	17.3%	17.7%	16.0%	17.3%
	Unifamiliar	14.3%	20.8%	17.9%	13.4%	12.6%	12.6%	17.3%	19.7%	19.2%

Fuente: elaboración propia

Aunque no son medidas totalmente comparables, el *NRMSE* y el *MAPE* guardan relación al ser medidas de error tipificadas, siendo la primera mucho más exigente ante errores extremos, y por tanto ofrece datos más altos, algo que se puede confirmar más adelante en la Tabla 5.19.

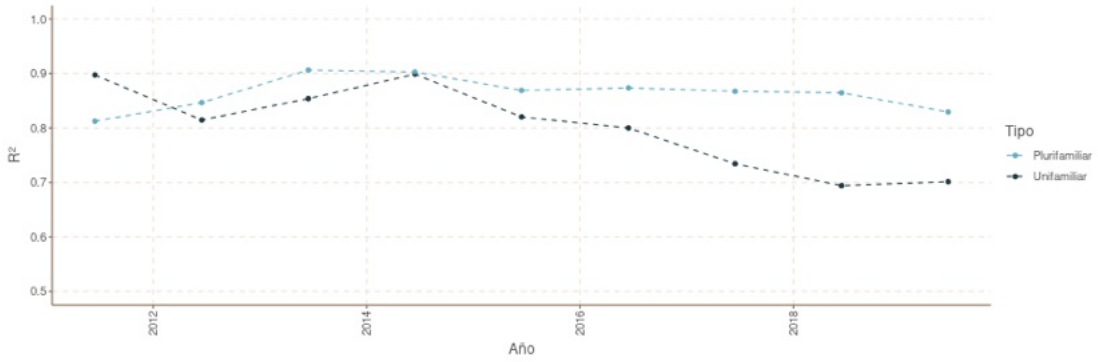
Cuando se compara el *NRMSE* de nuestro modelo con el *MAPE* del estudio de Alfaro *et al.* (2020), se observa como en nuestro caso los valores se encuentran en un orden de magnitud similar al *MAPE* medio municipal[8] de *Random Forests* para este estudio (15,93). Asimismo, este valor tampoco es muy diferente a los valores del modelo desarrollado por Clary y Lomax (2018), en cuyo caso, el *MEDAPE* es aún más optimista que el *MAPE*.

5.3.2.1 Selección de modelo: ensamblado o único

El modelo único muestra un comportamiento de ajuste decreciente, como se aprecia en la Figura 5.4, con una degradación más acusada en las viviendas unifamiliares a partir de 2015.

[8] Sobre 63 municipios españoles de más de 100.000 habitantes.

Figura 5.4. Niveles de ajuste del modelo único, desglosado por tipo



Fuente: elaboración propia.

El ajuste en las viviendas plurifamiliares se mantiene estable en el tiempor, como se aprecia en la Tabla 5.15.

Tabla 5.15. Ajuste de los modelos desglosados por tipologías (modelo único)

Tipo	Modelo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Unifamiliar	unico	89.7%	81.5%	85.4%	89.9%	82.0%	80.0%	73.4%	69.4%	70.1%
Plurifamiliar	unico	81.2%	84.6%	90.6%	90.3%	86.9%	87.4%	86.7%	86.5%	82.9%

Fuente: elaboración propia

El modelo único muestra errores cuadráticos menores en las viviendas plurifamiliares, con un mayor grado de estabilidad temporal en los errores, tal y como se ve en la Tabla 5.16. En cambio, para las viviendas unifamiliares, esta versión ofrece peores errores que el modelo agregado.

Tabla 5.16. RMSE (absoluto) y NRMSE (normalizado), modelo único

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Absoluta	Plurifamiliar	26,71	22,60	16,61	17,54	22,73	25,82	29,00	28,80	32,94
	Unifamiliar	16,16	21,33	17,40	13,52	18,85	20,63	24,36	29,55	30,26
Normalizada	Plurifamiliar	16.1%	14.4%	11.2%	11.8%	14.7%	15.3%	14.9%	13.4%	14.7%
	Unifamiliar	15.4%	21.2%	18.7%	14.9%	20.4%	21.1%	23.7%	26.8%	26.4%

Fuente: elaboración propia

La Tabla 5.17 compara la precisión del modelo ensamblado con respecto el único, a través de los errores y el R^2 en la muestra OOB. Se aprecia una notable diferencia en los resultados en función de la tipología. Por una parte, en las viviendas plurifamiliares el modelo único es mejor que el ensamblado en todos los periodos; y por otra, del caso de las unifamiliares, donde se produce lo contrario.

En términos de magnitud, también se aprecian diferencias, siendo las mejoras el doble de grandes en las plurifamiliares que en las unifamiliares.

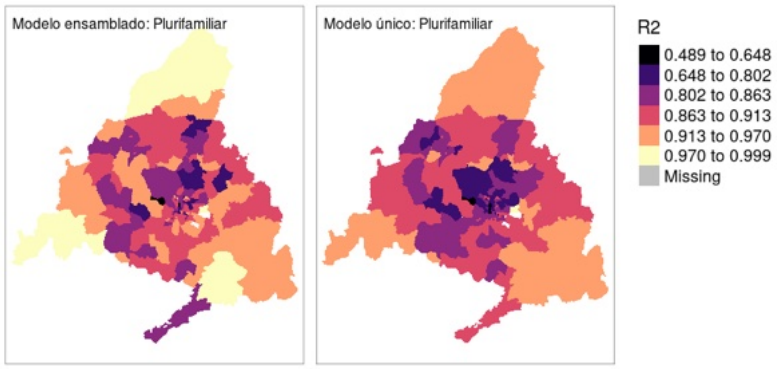
Tabla 5.17. Mejora (+) o empeoramiento (-) del modelo ensamblado, RMSE y R^2

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
RMSE	Plurifamiliar	-1,42	-1,99	-2,23	-2,51	-2,58	-3,36	-5,42	-5,74	-5,82
	Unifamiliar	1,16	0,45	0,80	1,33	7,24	8,27	6,58	7,85	8,22
R^2	Plurifamiliar	-2.0%	-2.8%	-2.7%	-3.0%	-3.1%	-3.5%	-5.4%	-5.9%	-6.6%
	Unifamiliar	1.4%	0.8%	1.3%	1.9%	10.2%	11.6%	10.5%	11.4%	11.4%

Fuente: elaboración propia

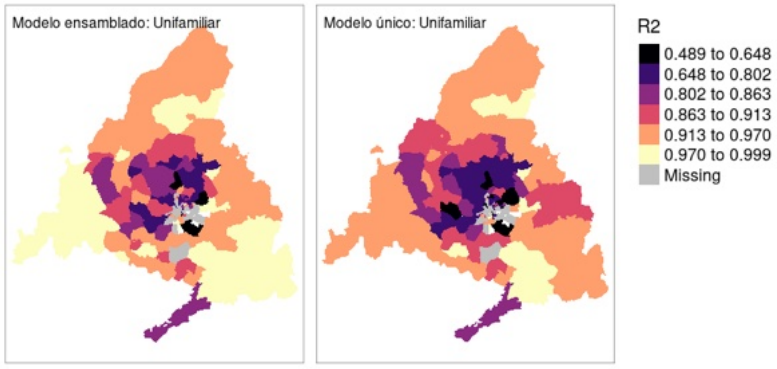
Dado que es deseable tener tanto un buen ajuste global como para las distintas zonas geográficas de análisis, se muestra el aspecto de ajuste zonal[9] en la Figura 5.6 y la Figura 5.5. Se aprecia que los modelos ensamblados mejoran al único resultados para todas las zonas, en los dos tipos de vivienda.

Figura 5.5. Ajuste espacial en vivienda plurifamiliar, Comunidad de Madrid



Fuente: elaboración propia.

Figura 5.6. Ajuste espacial en vivienda unifamiliar, Comunidad de Madrid

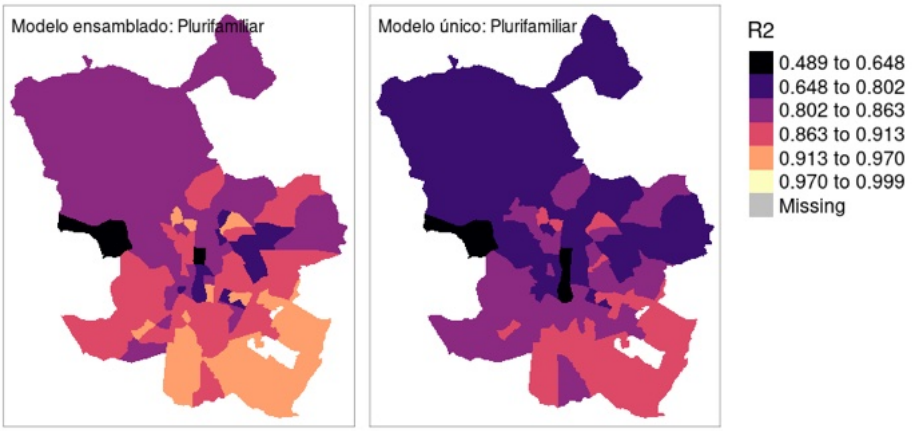


Fuente: elaboración propia.

[9] Datos calculados sobre todos los periodos desde 2011 a 2019, mediante validación cruzada (K=10).

Para el municipio de Madrid, se analiza sólo el ajuste para plurifamiliares, dada la poca representatividad de las unifamiliares. La Figura 5.7 se aprecia como los niveles de ajuste son menores en el modelo único, y con un alto grado de heterogeneidad zonal, siendo las regiones con mayor precisión las ubicadas en el sur y este de la ciudad. Por contra, las zonas con peor ajuste, son las el centro y el eje de la Castellana, debido a que en estas áreas existe una mayor heterogeneidad en precios y características.

Figura 5.7. Ajuste espacial en vivienda plurifamiliar, municipio de Madrid



Fuente: elaboración propia.

Desde un punto de vista numérico, los datos de las figuras anteriores se resumen en la Tabla 5.18, desglosados por tipo aproximación, tipo de vivienda y clase de zona. Se aprecia que los errores del modelo y coeficientes de determinación disminuyen siempre en los modelos ensamblados.

Tabla 5.18. Ajuste espacial promedio por clase de modelo y tipo

		Ensamblado		Único	
Capital	Tipo	R^2	RMSE	R^2	RMSE
No	Plurifamiliar	86,0%	1.29	79,9%	1.57
	Unifamiliar	80,6%	0.89	71,5%	1.22
Sí	Plurifamiliar	89,3%	0.67	87,1%	0.76
	Unifamiliar	91,3%	0.48	88,8%	0.59
Todos	Plurifamiliar	87,2%	1.06	82,5%	1.27
	Unifamiliar	85,9%	0.71	80,1%	0.94

Fuente: elaboración propia

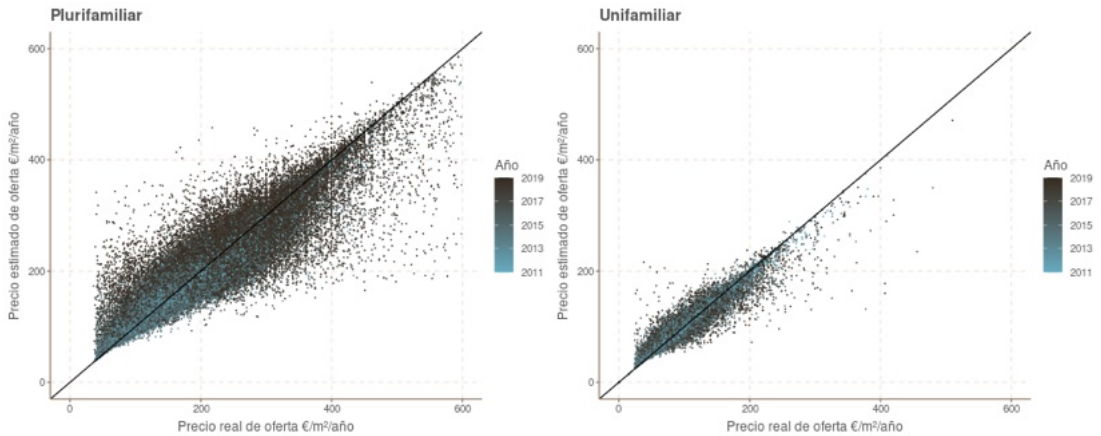
Los resultados globales y desglosados por zona contrapuestos se podrían

considerar un caso particular de la paradoja de Simpson[10] (Wagner, 1982). Esta discrepancia se debe a la ausencia de control de una variable de confusión relevante, que en este caso podría ser la geografía, cuya falta introduciría sesgos por la heterogeneidad espacial de los precios (McLaughlin y Young, 2018), o por un desigual nivel de información de los estratos que componen la muestra (Boeing, 2020). Ante esta divergencia de resultados, y debido a que el análisis geográfico al ser más completo potencialmente ofrece resultados más robustos (Steurer *et al.*, 2021), se opta por tomar el modelo que mejor funciona en este caso, es decir, el ensamblado.

5.3.2.2 Métricas sobre la población total (*in-bag*)

El desglose de las medidas, obtenidas sobre la población OOB es reducido, por una limitación técnica del algoritmo utilizado. Para disponer de un análisis en detalle, se han calculado una serie de métricas sobre la población completa (*in-bag*), que no pretenden estudiar la capacidad de generalización del modelo, sino las métricas clave, errores y sesgos de los diferentes estratos de la población. Un segundo motivo es que los precios estimados sobre el conjunto *in-bag*, se usan como base del índice de precios de oferta y como entrada del modelo de conversión definitivo. Por tanto, es necesario estudiar las diferencias entre los valores reales y los inferidos por el método. Para reducir sesgos de selección, en todas las medidas, se utiliza un remuestreo del tipo validación cruzada con 10 mezclas (K=10).

Figura 5.8. Ajuste de la regresión, precio de oferta



Fuente: elaboración propia.

La calidad del ajuste del modelo se representa en la Figura 5.8. La línea de 45º representa el ajuste perfecto, por tanto, cualquier desviación sobre ella indica un error de estimación del precio. Se aprecia que el segmento de viviendas

[10] La paradoja de Simpson se produce cuando el agregado de una medida sobre varias categorías muestra una incidencia mayor o menor al de cualquiera de las categorías individuales.

plurifamiliares tiene mayores desviaciones, especialmente los últimos años de la serie. Del mismo modo, los rangos de precios más altos son menos precisos, probablemente porque implican un mayor grado de heterogeneidad no controlada en el hedónico, bien debida a variables ausentes, o bien por falta de soporte de datos por ser un segmento muy minoritario.

La Tabla 5.19 muestra las medidas de error para viviendas plurifamiliares ponderadas por los factores de elevación. Se observa que el error cuadrático (RMSE) mantiene en órdenes de magnitud similares a los obtenidos en la métricas OOB, lo que indica que el modelo generaliza correctamente.

Tabla 5.19. Métricas in-bag de los modelos, viviendas plurifamiliares

Año	N	RMSE	NRMSE	MAE	MAPE	MEDAE	MEDAPE
2011	206.969	13.76	9,49%	7.72	5,32%	3.99	2,75%
2012	361.230	13.22	9,55%	7.40	5,35%	3.96	2,86%
2013	491.120	12.14	9,21%	6.63	5,03%	3.37	2,56%
2014	495.818	12.74	9,80%	6.64	5,11%	3.12	2,40%
2015	431.065	14.18	10,38%	7.10	5,20%	3.09	2,26%
2016	355.787	16.14	11,25%	8.06	5,62%	3.26	2,27%
2017	461.683	16.10	10,06%	8.00	5,00%	3.10	1,94%
2018	620.972	14.35	8,33%	7.01	4,07%	2.63	1,53%
2019	705.461	15.98	9,05%	7.34	4,16%	2.52	1,43%

Fuente: elaboración propia

Los errores porcentuales, tanto medios como medianos, son satisfactorios a la luz de lo comentado en el epígrafe 5.3.2. Además, se confirma que los errores cuadráticos normalizados son menores a los porcentuales absolutos, lo mismo que los valores medianos son siempre menores a los medios.

Los errores del segmento unifamiliar, recogidos en la Tabla 5.20, muestran valores muy bajos, aunque mayores que los presentados anteriormente para el conjunto OOB. Esto se puede atribuir a dos motivos: 1) las medidas OOB no se calculan ponderadas, y 2) que un menor tamaño poblacional y geográficamente más amplio, puede impactar en fenómeno de heterogeneidad espacial de los precios, que se sugería en el epígrafe 5.3.2.1, y mostrado gráficamente en la Figura 2.12 del capítulo 2.

Tabla 5.20. Métricas in-bag de los modelos, viviendas unifamiliares

Año	N	RMSE	NRMSE	MAE	MAPE	MEDAE	MEDAPE
2011	17.159	4.83	5,27%	1.99	2,18%	0.50	0,55%
2012	32.863	4.58	5,15%	1.82	2,05%	0.47	0,52%
2013	52.297	3.91	4,83%	1.41	1,74%	0.31	0,39%
2014	54.441	3.81	4,97%	1.30	1,70%	0.30	0,39%
2015	40.395	3.49	4,35%	1.15	1,44%	0.26	0,32%
2016	33.628	3.50	4,21%	1.16	1,40%	0.27	0,33%
2017	31.135	3.88	4,54%	1.13	1,32%	0.26	0,30%
2018	31.525	4.05	4,41%	1.18	1,29%	0.30	0,33%
2019	30.950	4.15	4,41%	1.19	1,26%	0.32	0,34%

Fuente: elaboración propia

Las métricas de error, desglosadas en función de si la vivienda se encuentra en la capital o el resto de la provincia, se recogen en la Tabla 5.21. Se observa que los errores en Madrid son más altos que los del resto de la Comunidad. Esta diferencia se debe a factores: el primero, que la proporción de viviendas unifamiliares es mayor en las zonas rurales y el extrarradio, como vemos en la Figura 2.12 del epígrafe 2.4.3; y el segundo, que la diversidad de zonas es mucho mayor en la ciudad, lo que implica una mayor dificultad en la especificación de los modelos, y por tanto, un mayor grado de errores de estimación.

Tabla 5.21. Métricas in-bag, capital o resto de provincia

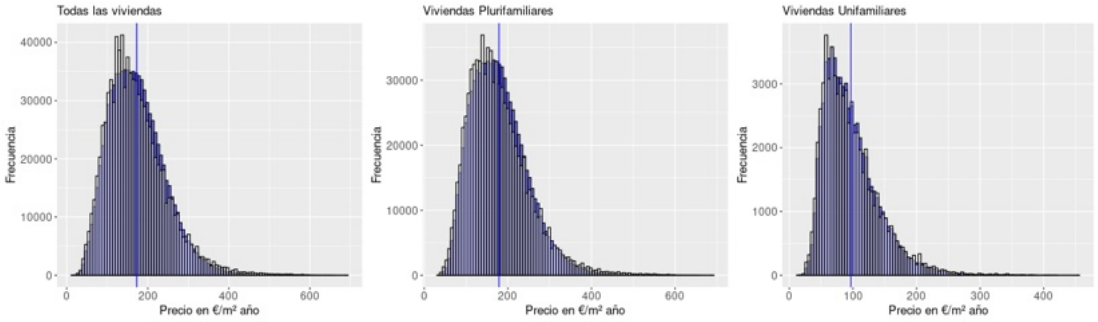
	Madrid						Resto					
Año	N	RMSE	MAE	MAPE	MEDAE	MEDAPE	N	RMSE	MAE	MAPE	MEDAE	MEDAPE
2011	159.508	16.27	9.97	6,02%	5.95	3,59%	64.620	7.62	3.73	3,44%	1.83	1,69%
2012	270.471	15.85	9.67	6,04%	5.88	3,67%	123.622	7.30	3.70	3,58%	1.91	1,85%
2013	358.096	15.29	9.33	5,99%	5.69	3,66%	185.321	5.87	3.01	3,04%	1.49	1,50%
2014	365.806	16.27	9.77	6,25%	5.81	3,72%	184.453	6.27	2.78	2,88%	1.22	1,27%
2015	319.303	18.81	11.02	6,62%	6.36	3,82%	152.157	6.04	2.73	2,68%	1.14	1,12%
2016	266.035	21.52	12.75	7,22%	7.25	4,10%	123.380	6.65	2.89	2,74%	1.15	1,09%
2017	374.100	21.33	12.64	6,48%	6.87	3,52%	118.718	6.22	2.65	2,26%	1.00	0,85%
2018	530.889	19.37	11.20	5,37%	5.85	2,80%	121.608	5.25	2.47	1,91%	0.97	0,75%
2019	601.289	20.45	11.14	5,24%	5.35	2,52%	135.122	9.35	3.37	2,46%	1.11	0,81%

Fuente: elaboración propia

5.3.3 Análisis de sesgos y residuos

Como se mencionaba en la metodología, los modelos ensamblados reducen de forma eficaz de reducir la varianza de los errores, pero pueden mostrar sesgos (Graczyk *et al.*, 2010). Se estudia el sesgo de los modelos a través de sus residuos, con una estrategia de remuestreo de tipo validación cruzada ponderada utilizando los pesos poblacionales. La Figura 5.9 muestra el sesgo de estimación a través de la función de densidad ponderada[11]. Se observa la existencia existe de un ligero sesgo positivo del error, que es mayor para las viviendas plurifamiliares.

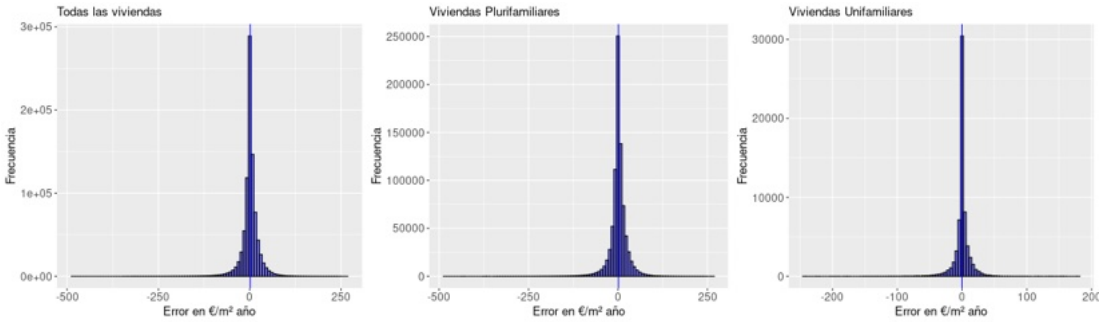
Figura 5.9. Histograma de precios reales (blanco) versus precios estimados por el modelo (azul), totales y desglosados por tipo de vivienda



Fuente: elaboración propia.

Los residuos muestran una alta concentración en torno al cero, como se observa en la Figura 5.10. Dichos errores se distribuyen de forma simétrica con respecto al origen, con un rango más amplio en los errores negativos que en los positivos. Por consiguiente, se deduce que los modelos tienden a infraestimar los casos más extremos, aunque ofrecen un buen comportamiento en el grueso de la población.

Figura 5.10. Distribución de estimación, totales y desglosados por tipo de vivienda



Fuente: elaboración propia.

Para analizar el grado de variabilidad de los errores, se ha calculado el coeficiente

[11] Se han utilizado los pesos poblacionales para ajustar la frecuencia de los valores en lugar de construir la función de densidad sobre las observaciones.

de dispersión ponderado (*COD*) sobre el cociente de la estimación sobre el valor real. Se toma la definición de la medida propuesta por Steurer *et al.* (2021) adaptada a una población ponderada, calculada según la siguiente expresión analítica:

$$COD = \frac{1}{W}\sum_{i=1}^{N}\left| w_i \frac{\hat{p_i}}{p_i} / \left(\sum_{i=1}^{N} w_i \left(\frac{\hat{p}}{p}\right)\right) - 1\right| \quad [5.7]$$

donde W es la suma de los pesos muestrales w_i, para cada observación i, con un precio real $\hat{p_i}$ y uno estimado $\hat{p_i}$.

En términos numéricos, resumidos en la Tabla 5.22, se observa que los errores de las viviendas de tipo plurifamiliar son positivos en todos los periodos, con un incremento gradual en el tiempo. Este comportamiento se puede generalizar para todo el espectro de valores, a la vista de los valores de los cortes cuantílicos y el rango intercuartílico (IQR)[12]. El COD, inferior al 5%, se encuentra en rangos muy inferiores a otras referencias como el de Steurer *et al.* (2021) con un 23,3%, o el de Alfaro (2020), con un valor medio de 15,56%, si bien, en ambos casos los modelos se referían precios de venta en vez de alquiler.

Tabla 5.22. Distribución error en €/m²/año, vivienda unifamiliar

Año	Media	P05	P25	P50	P75	P95	IQR	COD
2011	-0.44	-20.63	-4.05	0.03	4.18	18.34	8.23	2.9%
2012	0.05	-25.86	-6.02	0.42	7.32	25.23	13.34	5.0%
2013	-0.53	-19.03	-3.79	0.01	3.88	16.24	7.66	2.9%
2014	-0.69	-22.72	-4.09	0.04	4.22	18.68	8.32	2.8%
2015	-0.91	-27.27	-4.78	0.02	4.89	21.98	9.67	3.0%
2016	-1.14	-32.75	-5.96	0.03	5.95	26.65	11.91	3.2%
2017	-1.09	-40.42	-6.84	0.05	7.66	33.79	14.50	3.0%
2018	-0.84	-42.64	-6.51	0.04	7.51	37.08	14.02	2.4%
2019	-0.74	-49.22	-7.72	0.04	9.01	43.95	16.73	2.3%

Fuente: elaboración propia

En unifamiliares, los errores son mucho menores (aproximadamente la mitad que en plurifamiliar), y toman tanto valores negativos y positivos, como muestra la Tabla 5.23. La una distribución es relativamente simétrica (en función de la diferencia entre el P05 y el P95), con una diferencias de signo entre la mediana y la media, que indica que el modelo es más impreciso en las predicciones de los precios altos más extremos. El coeficiente de dispersión, también se encuentra en valores muy bajos y estables en el tiempo.

[12] El rango intercuartílico se calcula como la diferencia en valor absoluto del valor del percentil 75 y el percentil 25.

Tabla 5.23. Distribución error en €/m²/año, vivienda unifamiliar

Año	Media	P05	P25	P50	P75	P95	IQR	COD
2011	0.20	-9.94	-1.72	0.07	2.06	11.04	3.78	2.3%
2012	0.08	-9.25	-1.59	0.03	2.09	10.05	3.68	1.9%
2013	0.04	-10.59	-1.66	0.03	1.97	10.74	3.63	1.8%
2014	0.15	-12.80	-1.63	0.04	2.30	13.77	3.93	1.8%
2015	0.21	-14.07	-1.77	0.00	2.42	15.91	4.19	1.4%
2016	0.21	-15.96	-1.87	0.04	2.83	15.66	4.70	1.4%
2017	-0.43	-28.18	-3.59	0.06	5.55	22.88	9.14	1.4%
2018	-0.13	-33.98	-4.12	0.24	6.78	29.02	10.90	1.4%
2019	-1.13	-33.77	-5.17	0.08	6.05	26.27	11.22	1.3%

Fuente: elaboración propia

5.3.4 Interpretabilidad, importancia de variables

Hasta ahora se ha analizado el funcionamiento del modelo en términos de ajuste, pero no se ha entrado en la contribución de las distintas variables, que es la base en la que se sustenta la teoría de precios hedónicos. Aunque los modelos de árboles han sido denominados históricamente como “cajas negras”, porque no generan coeficientes para cada característica (Rico y Taltavull, 2021), existen mecanismos que permiten interpretar las contribución en términos de importancia[13]. De forma general, hay 3 formas de evaluar la importancia de las características, con respecto a los poderes predictivos del modelo:

- De filtro: las características se filtran independientemente del modelo a través de criterios en sus propias propiedades (correlación con el objetivo
- De envoltorio (*wrapper*): basados en algoritmos de búsqueda que tratan a los predictores como entradas y utilizan el error como la salida a optimizar.
- Integradas: la selección de características integradas combinan algoritmos de búsqueda del predictor con la estimación de parámetros y, generalmente, se optimizan con una única función objetivo.

La estimación de la importancia de las variables en un modelo de ensamblado se debe realizar de forma diferente a la de un modelo de regresión por mínimos cuadrados, como describe Grömping (2009). Para el caso de *Random Forests*, los dos métodos más comunes son el de impureza y el de permutación. El de impureza mide la contribución de la variable en la reducción de la medida de desorden del modelo. El de permutación mide el impacto que tendría en la precisión si se permutan los valores entre todos los registros de la muestra[14].

[13] La importancia de cada variable es una medida que relaciona la capacidad marginal de cada independiente de explicar la variable dependiente.

[14] Se mide el efecto de usar el valor real, sobre un valor aleatorio que sigue la misma distribución

La forma más común de determinar la importancia de las variables es la impureza (Strobl *et al.*, 2007), basada en el criterio usado internamente para la determinación de los cortes del árbol de decisión. Cada uno de los cortes son una condición que se aplica a una sola característica (ejemplo: área mayor de 100 metros), y subdivide el conjunto de datos en dos grupos, de forma que la entropía de la variable respuesta disminuye en ambos subconjuntos, o dicho de otro modo, divide la población en los dos grupos lo más homogéneos posible. Para estimar el criterio de corte, se elige la condición sobre una variable localmente[15] la reducción de la varianza, la cual es mayor para los precios de cualquiera de los grupos resultantes que para el conjunto original.

Puesto que se conocen el impacto que tiene el corte sobre una variable, en la reducción de impureza para cada uno de los cortes, la disminución de impurezas de cada característica se calcularía como el un promedio ponderado de estas medidas.

La importancia se ha calculado de forma general para todo el modelo, por tanto, no es posible determinar las contribuciones marginales para cada registro predicho. No obstante, existen métodos más sofisticados, como los basados en Shapley (Roth, 1988), que aproximan la importancia de una solución según el reparto de coste y beneficios entre los colaboradores, en este caso los predictores. Dicho de otra manera, la esperanza de la contribución marginal de cada predictor, para todos los casos posibles (Winter, 2002). Este método ha evolucionado en los últimos años en un marco que une la idea original y la teoría de juegos, denominado SHAP[16] Lundberg y Lee (2017) y Lundberg *et al.* (2018) desarrollan en profundidad los últimos avances en este método aplicado a ensamblados de árboles. A modo de ejemplo, Rico y Taltavull (2021) desarrollan un análisis en profundidad de los valores SHAP aplicados a un modelo de precios de la vivienda en Alicante.

En este caso, la medida de importancia se ha tipificado a valores entre 0 a 100, en la que 100 representa el mayor grado de importancia por variable. Puesto que el cálculo de la importancia de las variables para el modelo ensamblado es complejo de intepretar, se estudiarán las contribuciones de las variables para cada una de las partes del ensamblado.

Para el modelo de vivienda plurifamiliar, los atributos de superficie, número de habitaciones tamaño y calidad del edificio son los más relevantes, como se ver en la Tabla 5.24. A pesar de que las medidas son difícilmente comparables

[15] Se indica que es localmente porque la evaluación del impacto sobre la entropía se realiza para ese corte no para la población general del árbol.

[16] SHapley Additive exPlanations (SHAP) es un método que permite calcular, a nivel de observación individual, el grado de contribución de las distintas variables de entrada en el resultado inferido por un modelo de aprendizaje automático.

de estudio en estudio, este resultado es consistente con la presencia en las primeras posiciones en importancia de las variables de otros trabajos, dando mayor prioridad a aquellas que contienen la superficie, antigüedad o número de habitaciones (Clark y Lomax, 2018; Füss y Koller, 2016; Hanink *et al.*, 2012; Rico y Taltavull, 2021).

Tabla 5.24. Importancia de variables plurifamiliar modelo de atributos

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Num. De habitaciones	100.0%	12	Tiene terrazas	13.2%
2	Año de construcción	99.7%	13	El piso tiene anejos	11.9%
3	Número de inmuebles en finca	56.1%	14	Canal de venta del inmueble	8.9%
4	Calidad de la construcción	48.1%	15	Es un estudio	8.4%
5	Número de pisos en edificio	44.0%	16	Tiene armarios empotrados	7.6%
6	Planta en el edificio	24.9%	17	Posición vertical en el edificio	7.6%
7	Tiene ascensor	16.2%	18	¿Tiene su edificio una piscina?	7.1%
8	Tipo de instalaciones de la finca	15.6%	19	Es un dúplex	6.5%
9	¿Interior o exterior?	15.0%	20	Es un ático	4.1%
10	¿Tiene aire acondicionado?	14.2%	21	Nuevo o segunda mano	3.4%
11	Periodo en Año-Mes	13.8%	22	Tiene balcón	1.3%

Fuente: elaboración propia

Tabla 5.25. Importancia de variables plurifamiliar modelo de utilidad

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Superficie total construida	100.0%	15	Accesibilidad WALK 6	14.3%
2	Población del municipio	61.6%	16	Accesibilidad CAR 4	14.3%
3	Accesibilidad CAR 2	44.6%	17	Accesibilidad WALK 7	14.0%
4	Densidad de población	34.4%	18	Accesibilidad WALK 8	13.6%
5	Proporción alquiler/venta	33.4%	19	Accesibilidad WALK 9	12.9%
6	Num. de inmuebles en alquiler	25.6%	20	Pct. Estudios estudios superiores	12.7%
7	Accesibilidad WALK 1	24.5%	21	Accesibilidad WALK 3	12.4%
8	Número de contactos en la zona	23.1%	22	Accesibilidad CAR 9	11.9%
9	Accesibilidad WALK 2	20.9%	23	Accesibilidad CAR 3	11.6%
10	Accesibilidad WALK 4	20.1%	24	Accesibilidad CAR 6	11.3%
11	Accesibilidad CAR 5	19.3%	25	Accesibilidad CAR 7	10.9%
12	Accesibilidad CAR 8	18.1%	26	Porcentaje de mayores	10.7%
13	Accesibilidad CAR 1	15.5%	27	Tasa de extranjeros	9.8%
14	Accesibilidad WALK 5	14.4%	28	Inmuebles en venta en la zona	7.8%

Fuente: elaboración propia

El modelo utilidad combina los atributos constructivos básicos con los de localización y los de dinámicas del mercado (*LEADS*, *RENT_SALE_RATIO* y *LEADS_RESIDENTIAL*), como se aprecia en la Tabla 5.25. Todos atributos

accesibilidad mantienen una importancia similar, destacando ligeramente aquellas procedentes del medio de transporte a pie. Como sucede en las múltiples referencias de la literatura, no existe una gran diferencia entre la importancia de las distintas medidas de accesibilidad, repartiéndose la importancia casi de forma equitativa, con la tendencia de aparecer ciertas medidas demográficas como la densidad de población en las primeras posiciones, como en (Rico y Taltavull, 2021).

Tabla 5.26. Importancia de variables unifamiliar, modelo de atributos

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Inmuebles en alquiler	100.0%	18	Canal de comercialización	7.8%
2	Superficie construida	93.5%	19	¿Tiene su edificio una piscina?	7.7%
3	Área útil	90.8%	20	Periodo Año-mes	7.6%
4	Inmuebles en venta	59.8%	21	Tiene armarios empotrados	6.5%
5	Proporción alquiler/compra	58.5%	22	¿Tiene aire acondicionado?	6.2%
6	Tipo de zona (cluster)	57.2%	23	Tiene jardín	6.0%
7	Contactos medios en zona	40.8%	24	Tamaño del garaje	5.6%
8	Año de construcción	30.1%	25	Tiene almacenamiento / Trastero	4.8%
9	Tamaño de la parcela	26.0%	26	Está orientado al sur	4.3%
10	Calidad de la construcción	22.9%	27	Tiene terrazas	4.0%
11	Número de habitaciones	20.4%	28	Está orientado al norte	3.6%
12	Número de dormitorios	20.3%	29	Tiene portero	3.6%
13	Número de baños	20.1%	30	Está orientado al este	3.1%
14	Número de pisos en edificio	10.9%	31	Nuevo o segunda mano	2.7%
15	Tipo de inmueble unifamiliar	10.5%	32	Está orientado al oeste	2.6%
16	Certificado energético	8.9%	33	Tipo de garaje	0.1%
17	Tipo de instalaciones en finca	8.6%			

Fuente: elaboración propia

Para el modelo de atributos de unifamiliar, los atributos de dinámicas de mercado junto con los constructivos y el clúster de zonas, acaparan la mayor parte de la reducción de la impureza, como se ver en la Tabla 5.26.

En el modelo de utilidad unifamiliar, como se aprecia en la Tabla 5.27, la importancia relativa de las medidas de accesibilidad es mucho mayor que en el caso de plurifamiliar. Por otra parte, la variable más relevante es la población del municipio (en lugar de la superficie).

Tabla 5.27. Importancia de variables plurifamiliar, modelo de utilidad

Pos.	Variable	Imp.	Pos.	Variable	Imp.
1	Población del municipio	100.0%	13	Accesibilidad WALK 7	46.2%
2	Accesibilidad CAR 2	84.1%	14	Accesibilidad CAR 3	46.1%
3	Accesibilidad CAR 5	73.0%	15	Accesibilidad WALK 2	45.7%
4	Accesibilidad WALK 6	70.4%	16	Accesibilidad CAR 9	45.3%
5	Pct. personas con estudios superiores	65.3%	17	Accesibilidad WALK 9	44.8%
6	Densidad de población	63.0%	18	Accesibilidad CAR 7	43.8%
7	Accesibilidad CAR 8	62.1%	19	Accesibilidad WALK 8	41.0%
8	Accesibilidad WALK 1	54.3%	20	Accesibilidad CAR 6	40.2%
9	Accesibilidad WALK 5	53.9%	21	Accesibilidad WALK 3	38.5%
10	Accesibilidad WALK 4	53.3%	22	Porcentaje de mayores	33.4%
11	Accesibilidad CAR 4	52.3%	23	Tasa de extranjeros	28.8%
12	Accesibilidad CAR 1	47.4%			

Fuente: elaboración propia

Estos resultados guardan consistencia con otros estudios[17], pero sería interesante desarrollar un análisis más en profundidad, ya que como demuestran empíricamente Rico y Taltavull (2021), la influencia de las variables difiere entre los estratos de la muestra.

El método propuesto para construir los de modelos hedónicos de oferta en el presente Capítulo, cuenta con un alto grado de desglose y permite introducir una mayor precisión a las estimaciones del modelo de mercado, presentado en el Capítulo 3. Sin embargo, la aplicación de estas estimaciones requieren un ajuste del planteamiento original del modelo de conversión oferta-mercado, debido a los sesgos derivados de la ausencia de control de la localización. En el próximo capítulo se presenta un método que corrige este efecto, y permite calcular estimaciones del precio de mercado con un alto grado de detalle.

En este capítulo se ha desarrollado un modelo que estima,de forma precisa los precios de oferta, y que aprovecha la contribución de varios enfoques para crear un estimador más robusto. En el siguiente capítulo, se presentará una metodología capaz de corregir los sesgos zonales de los que adolece el modelo hedónico de mercado presentado en el capítulo 3.

[17] Como por ejemplo, Rico y Taltavull (2021), Füss y Koller (2016) o Ĉeh (2018).

Anexo 5a. Clasificación automática de zonas

Como paso previo a la construcción del modelo ensamblado, se crea una variable sintética que pueda relacionar los distintos modelos atendiendo a la diversa naturaleza de las zonas. Este tipo de zona (denominado como *Cluster* en los modelos) que representa una categorización natural de secciones censales. Esta categorización se desarrollada mediante el método de aprendizaje automático no supervisado K Medoides (Schubert y Rousseeuw, 2019), que es una versión robusta del algoritmo de análisis clúster K Medias (Lloyd, 1982). Para evitar distorsiones de escala y colinealidad se preprocesan las variables mediante un proceso de escalado multimensional (SMACOF) a través de un matriz simétrica de disimilitud (Borg y Groenen, 2005).

Se trabaja en dos ámbitos: el primero para Madrid, y el segundo, para el resto de la Comunidad, con 7 grupos distintos en cada ámbito (K=7). Las fuentes de información son datos los de renta y viviendas de la Comunidad de Madrid [18], puntos de interés de Open Street Map (2017).

Para cada sección censal de la capital, se usan datos que recogen las características zonales de tipo demográfico, económico, y de oferta de servicios de turismo y ocio, que se reduce a las siguientes variables:

- Número de viviendas.
- Número de comercios.
- Número de bares-restaurantes.
- Número de hoteles.
- Número de museos.
- Número de monumentos.
- Número de negocios de tipo industria.
- Paradas de transporte público.
- Nivel educativo.
- Porcentaje de población extranjera.
- Número de centros sanitarios
- Renta media por persona.

Mientras que para el resto de provincia, se parte de la misma base, excluyendo las de tipo turístico:

- Número de viviendas.
- Número de comercios.
- Número de bares-restaurantes.
- Número de hoteles.

[18] Portal estadístico de la Comunidad de Madrid (https://www.madrid.org/iestadis).

- Número de negocios de tipo industria.
- Renta media por persona.

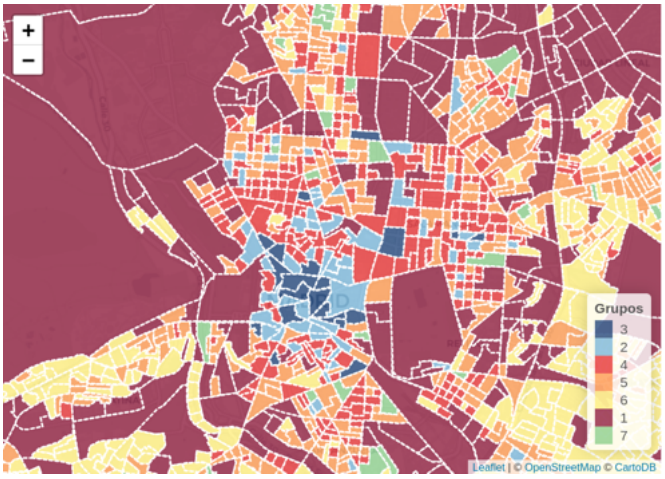
Se generan 7 grupos de zona, numerados del 1 al 7, y descritos en la Tabla 5.28.

Tabla 5.28. Descripción de grupos para la ciudad de Madrid

Código	Descripción	Secciones censales
1	Zona con la renta más alta. Muy baja densidad de viviendas, comercios y bares-restaurantes	471
2	Gran densidad de viviendas y lugares de ocio. Alta densidad de comercios y bares-restaurantes	66
3	Alta densidad de viviendas. Gran densidad de comercios, bares-restaurantes, hoteles y monumentos	21
4	Gran densidad de viviendas. Alta densidad de comercios y bares-restaurantes	313
5	Media densidad de viviendas, comercios y bares-restaurantes	587
6	Zona con la renta más baja. Baja densidad de viviendas, comercios y bares-restaurantes	897
7	Zona industrial. Baja densidad de viviendas. Media densidad de comercios y bares-restaurantes	51

Fuente: elaboración propia

Figura 5.11. Clusters de zona en el municipio de Madrid



Fuente: elaboración propia.

En el centro (véase Figura 5.11), dónde ubica la actividad turística y de ocio, se produce una alta concentración de los grupos 2 y 3; los grupos 4 y 5 se extienden desde el centro hacia el norte y el sur por el eje Prado-Castellana (división norte

sur de la ciudad); los grupos 6 y 1 se ubican en el anillo exterior, con mayor presencia del grupo 1 en el norte y oeste; y por último, el grupo 7 solamente está presente en el distrito de Villaverde (sur).

La agrupación zonal para el resto de la Comunidad[19], recogida en la Tabla 5.29, distingue principalmente las área rurales y las áreas residenciales de rentas altas. En las últimas, predomina la vivienda de tipo unifamiliar y se concentran en la parte norte-oeste del anillo de municipios inmediatamente exterior a la capital.

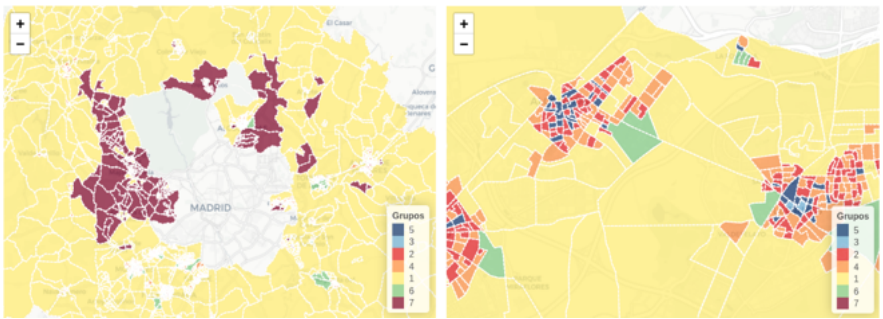
Tabla 5.29. Descripción de grupos resto de la Comunidad de Madrid

Código	Descripción	Secciones censales
1	Muy baja densidad de viviendas, comercios y bares-restaurantes	652
2	Alta densidad de viviendas, comercios y bares-restaurantes	371
3	Alta densidad de viviendas, comercios, bares-restaurantes y hoteles	18
4	Media densidad de viviendas, comercios y bares-restaurantes	510
5	Gran densidad de viviendas, comercios y bares-restaurantes	107
6	Zona industrial. Baja densidad de viviendas. Densidad de comercios y bares-restaurantes	30
7	Zona con la renta más alta. Baja densidad de viviendas, comercios y bares-restaurantes	175

Fuente: elaboración propia

Estas últimas agrupaciones tienen un menor nivel de variabilidad, como se observa la Figura 5.12.

Figura 5.12. Cluster resto de la Comunidad de Madrid



Fuente: elaboración propia.

[19]Aunque las zonas compartan los códigos numéricos del clustering de la ciudad se refieren a tipos de zonas distintas.

Capítulo 6

Modelo hedónico final

"El conocimiento no sirve para nada a menos que se ponga en práctica"

— Anton Chejov

6.1 Introducción

Los capítulos anteriores se han centrado en la construcción de un modelo preciso que represente de forma fiel el comportamiento del mercado, que replique los valores de las estadísticas oficiales disponibles. Asimismo, se estima una función que transforma los precios de oferta en precios reales de alquiler. El modelo de oferta, por otra parte, permite estimar los precios en comercialización según un conjunto muy extenso de características, aportando herramientas precisas de estudio de los factores que contribuyen al valor de mercado de las viviendas.

El método anterior produce series temporales de las rentas con un alto grado de descomposición, reduciendo los sesgos de omisión de variables de oferta que existen con el uso de los datos originales de fuentes oficiales. Sin embargo, existe aún un inconveniente a resolver en esta aproximación, identificado en el apartado 3.4.3 del Capítulo 3, que es la dificultad del modelo de mercado de reproducir los comportamientos específicos de cada zona. Esta cuestión está motivada, principalmente, por la ausencia de información geográfica específica en los datos de la EPF principalmente, y para solucionarlo requiere de una corrección zonal de precios.

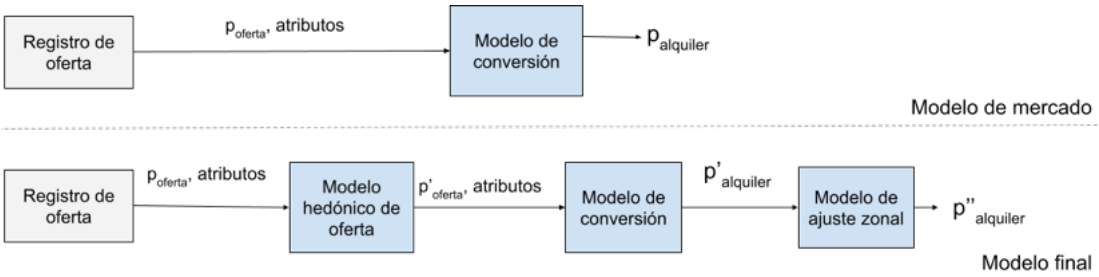
Este capítulo presenta un modelo hedónico denominado "final", cuyo objetivo es corregir los desequilibrios zonales de los métodos usados en los capítulos anteriores. Para ello, se utilizan las series de precios anuales de alquiler del "Sistema Estatal de Índices de Referencia del Precio del Alquiler de Vivienda",

publicadas por el MITMA (2020), y cuya utilidad se apoya en una alta correlación entre estos datos y los de oferta (véase (Rey-Blanco *et al.*, 2023a)), tanto en sus valores absolutos como en sus variaciones interanuales.

La cointegración entre el dato de oferta y el final ha sido ampliamente analizada en la literatura, y se deriva del proceso secuencial en el que se forman los distintos precios (Shimizu *et al.*, 2016). De igual manera, significar estudios como el de Kokot (2015), sobre precios de alquiler en Polonia, o el de Ardila *et al.* (2021) para el mercado suizo.

El proceso de cálculo del modelo final se realiza en los tres pasos descritos en la Figura 6.1. El flujo superior, muestra el cálculo del precio de alquiler por el modelo de mercado original, $p_{alquiler}$; la secuencia de la fila inferior, representa el proceso completo del modelo final, a través del encadenado de tres modelos: el hedónico de oferta, el modelo de conversión a precios de mercado y el modelo de ajuste zonal, con el resultado final $p''_{alquiler}$.

Figura 6.1. Proceso de estimación del precio del alquiler usando precio de oferta del modelo hedónico



Fuente: elaboración propia.

El capítulo se estructura en dos partes: la primera, contiene una descripción metodológica del proceso de corrección de sesgo zonal, y la segunda, que analiza los resultados obtenidos desde las ópticas de reducción del sesgo zonal, sesgo general de los precios, y validez del modelo para proyectarse a periodos futuros.

6.2 Metodología

El sesgo identificado en el apartado 3.4.3 del Capítulo 3, procede de la incapacidad de controlar la heterogeneidad espacial en los precios de la vivienda con los atributos de los datos de la EPF. Esta heterogeneidad, significa que la lógica sobre que se basa la formación de los precios no es un proceso estacionario espacialmente, por tanto, la relación funcional entre las covariables y el precio tiene peculiaridades diferentes en función de la unidad geográfica de análisis. En el Anexo I del presente capítulo, se describe en mayor profundidad las cuestiones de identificación y corrección de sesgos aplicadas en la metodología.

El fenómeno de la heterogeneidad espacial está ampliamente documentado y se analiza con profundidad por Hu (2022), Wu (2020), Helbich (2014), Páez (2008) y Kestens (2006). En nuestro caso, sus consecuencias se resumen en que el modelo, al no disponer de referencia a las zonas concretas, tiende a representar los precios de las zonas como un compuesto de las medias del valor de un estrato macro zonal (por ejemplo: zona periférica de altos ingresos o municipio de más de 50.000 habitantes), de lo que se deriva la imposibilidad de controlar la variabilidad de precios, como se verá más adelante en los ejemplos de las Figuras 6.2.

La ausencia de control por la zona geográfica provoca una asignación de precios arbitraria ante un desglose de datos geográfico-funcional, puesto que el modelo solo es capaz de asignar precios a través de la dimensión funcional. Como consecuencia, se producen dos fenómenos: 1) el dato desglosado por zona muestra una alta irregularidad en el tiempo[1], y 2) al calcular agregados con más de una dimensión de estratificación, el resultado también muestra irregularidades.

Figura 6.2. Precios de mercado, oferta y precios MITMA, vivienda plurifamiliar

Fuente: elaboración propia.

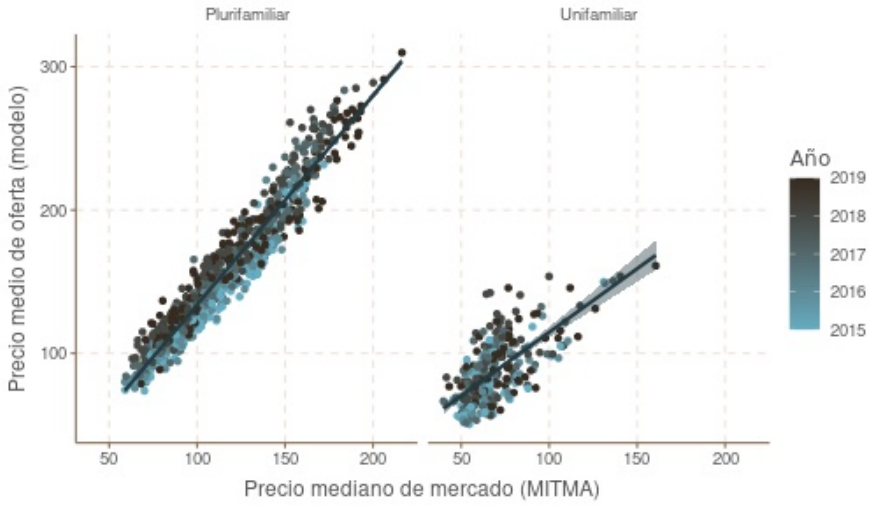
Para demostrar esta situación gráficamente y a modo de ejemplo, se representan en la Figura 6.2, los precios de mercado, oferta y de MITMA para tres zonas

[1]El término "irregularidad" se refiere a secuencias de precios incoherentes con respecto a la tendencia general del mercado, por ejemplo inconsistencias en la evolución del precio de mercado con respecto al precio de oferta.

diferenciadas. Para el caso de Mejorada del Campo[2], se observa que la serie de oferta y la de MITMA están cointegradas, mientras que la serie del modelo de mercado tiene una tendencia totalmente distinta a partir de 2016. En el caso de El Plantío-Aravaca[3], se aprecian diferencias en las pendientes de las series y una gran diferencia entre el precio medio del modelo y el registrado por MITMA . El tercer ejemplo, para San Martín de la Vega[4] muestra una anomalía en la serie de oferta y mercado para el año 2019, atribuible a anomalías del precio de ofert a.

Una propiedad muy interesante de estas series de precios es la alta correlación entre los precios de mercado y de oferta. A este respecto, existen varios artículos que analizan el nivel de relación entre las poblaciones de oferta y de transacción: Chapelle *et al.* (2022), para Francia, y Kokot (2015), en Polonia, estudian el nivel de correlación de transacciones y oferta obteniendo un coeficiente de correlación de 0,95 y 0,99 respectivamente.

Figura 6.3. Correlación entre precio medio de oferta y mediano del MITMA



Fuente: elaboración propia.

La correlación entre precios es mucho más fuerte en viviendas plurifamiliares que en unifamiliares, como se aprecia en la Figura 6.3. Para las últimas, la variabilidad aumenta a medida que se incrementan los precios del mercado. Por otra parte, se observa una relación cambiante en el tiempo, atribuible a las condiciones coyunturales del mercado, por ejemplo, la pendiente menos pronunciada del 2015 puede indicar una fase de mayor rotación de los alquileres. Existen distintas referencias bibliográficas que demuestran que la relación precios de oferta/mercado no son inmutables, como la aportación de Han (2016),

[2] Se toma este municipio por ser un área con una muestra pequeña e irregular.
[3] Se toma la zona por estar en el área metropolitana cercana a Madrid y tener una amplia muestra.
[4] Se selecciona esta zona por ser un área rural lejana a la capital.

quien muestra como se produjeron variaciones de entre el 15 y 30%, durante el boom inmobiliario de los años 2000-2007 en Estados Unidos.

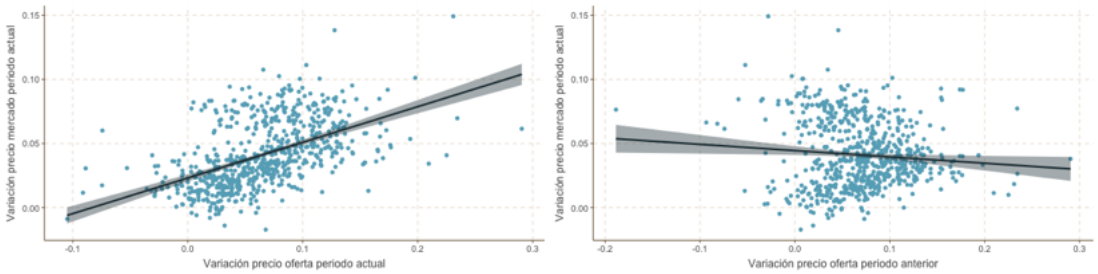
Las variabilidad de la Figura 6.3 se puede, además, justificar por la diferente de composición de ambas poblaciones[5], aún cuando los valores medios de las poblaciones estén fuertemente correladas. A este respecto, Shimizu (2016), Kolbe (2021) y Ardila *et al.* (2021) han constatado que las distribuciones de frecuencias de precios oferta y de mercado no son iguales, y cuyas diferencias se pueden deber a una sobrerrepresentación en oferta de los inmuebles menos líquidos y la infrarrepresentación de los más líquidos, o también, a la presencia de anuncios en régimen de subasta (Han y Strange, 2016). Por otra parte, Ardila *et al*. (2021) destacan que el dato de oferta tiende a infraestimar la magnitud de los cambios de tendencia.

Por otra parte, Diaz y Jerez (2013) argumentan que las fricciones en el proceso de búsqueda produce retrasos en dicho proceso y, consecuentemente, introduce volatilidad en los precios. Estas fricciones están presentes en manera desigual en los submercados que componen la muestra por los desequilibrios entre la oferta o la demanda, por lo que la única forma de controlarlas en los modelos es mediante el uso de variables que representen las características de cada mercado.

La cointegración entre series de alquiler y oferta también ha sido documentada en la literatura. Por ejemplo, Kokot (2015) comprueba la existencia de cointegración más un retraso temporal de 5 meses entre la serie de mercado y la de oferta. En nuestro caso, se han analizado la relación entre las variaciones anuales del precio de alquiler y las de oferta, identificando que las variaciones en los precios de mercado están correlacionadas con la variación del precio de oferta del año anterior, como se puede comprobar gráficamente en la Figura 6.4. Esta relación no es tan fuerte como la que existe entre los precios, porque en este caso el coeficiente de correlación de Pearson entre las variaciones es de 0,51. Este menor grado de relación podría estar motivada por dos cuestiones: 1) que el periodo de retraso entre series sea menor de un año, como el caso de (Kokot y Bas, 2015), y 2) que esta relación pueda variar en función de la zona. El segundo argumento se sustenta sobre la dependencia de la ratio de precios y la intensidad de la demanda de la zona (Han y Strange, 2016; Han y Strange, 2014; Shimizu *et al.*, 2016). Por tanto, como el modelo no incorpora estos efectos en un mercado heterogéneo, es plausible la existencia de variabilidad no controlada por este motivo.

[5]En el caso de la población de MITMA se desconocen.

Figura 6.4. Relación variación precio de mercado y oferta (vivienda plurifamiliar)



Fuente: elaboración propia.

Sobre la base anterior, se construyen dos modelos de regresión por mínimos cuadrados ordinarios, uno para cada tipologías de vivienda. Las variables independientes son la variación de oferta con uno y dos periodos de retraso. Para el caso de el tipo plurifamiliar, los coeficientes del modelo se recogen en la Tabla 6.1, y se observa como las variaciones de precios son altamente representativas. Los modelos permiten reconstruir la variación del precio de mercado a partir del histórico de variaciones del precio de oferta. Lo que desde un punto de vista económico representa la capacidad de la zona para absorber los inmuebles en oferta, de tal forma que se sustituyen inmuebles en oferta por inmuebles de mercado, produciéndose una transferencia del precio de oferta sobre el de mercado.

Tabla 6.1. Coeficientes del modelo de absorción - viviendas plurifamiliares

	Estimación	std.error	t value	Pr(>|t|)	signif.
(Intercept)	0.03	0.00	16.12	< 2e-16	***
VAR_ASKING	-0.07	0.02	-4.57	6.10e-06	***
VAR_ASKING_1	0.30	0.02	19.55	< 2e-16	***

Códigos signif.: *** 0,001 ** 0,01 * 0,05 . 0,1 1
Error estándar de los residuos: 0.0169 sobre 564 grados de libertad (DF)
R^2: 0,423, R^2 ajustado: 0,421
F-statistic: 2017 sobre 2 y 564 DF, p-value: < 2,2e-16
Num. observaciones: 564

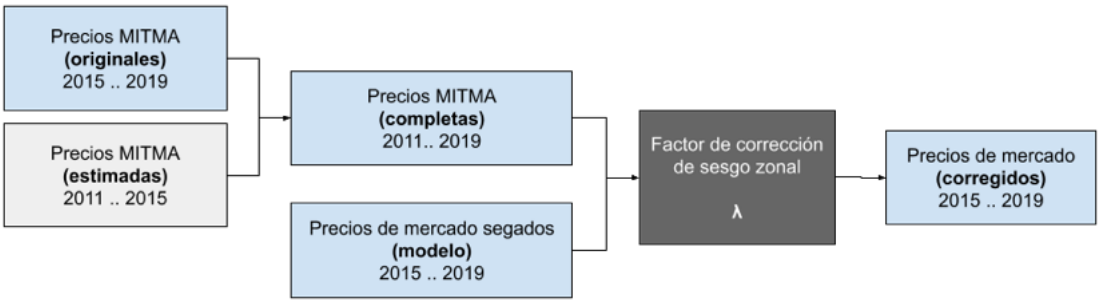
El coeficiente de determinación de la regresión anterior indica que el modelo no explica totalmente la varianza de la variable objetivo. Es evidente que parte de esta varianza es explicada mediante la introducción de nuevas variables, como el nivel de demanda de la zona o la evolución de volumen inmuebles en oferta (De Wit *et al.*, 2013). No obstante, el uso de series de precios anuales provoca que la varianza debida a retrasos menores a 12 meses sean difícilmente controlables.

La metodología se basa en la estimación de una serie de ratios simples[6] sobre un una muestra estratificada (Lohr, 2019), aplicada en dos etapas. El proceso crea, para cada estrato e (definido como la combinación entre zona, tipo de vivienda y un año t), un factor dinámico de corrección de sesgo del precio, denominado en adelante $\lambda_{t,e}$, y que se aplica para calcular los nuevos precios de mercado $\hat{P}$:

$$\hat{P}_{t,e} = \lambda_{t,e} \times P_{t,e} \quad [6.1]$$

El proceso se describe, de forma general, en la Figura 6.5.

Figura 6.5. Pasos del proceso de corrección del sesgo zonal



Fuente: elaboración propia.

Los factores a aplicar a los precios se calculan sobre los datos MITMA disponibles[7]. Los factores λ relacionan el precio medio ponderado con la mediana del precio, para el estrato de la muestra conocida, P^M. Este factor de ajuste, denominado $\lambda^M_{t,e}$, se calcula a través de la expresión:

$$\lambda^M_{t,e} = \frac{P_{t,e}}{P^M_{t,e}}, \forall t \in [2015..2019] \quad [6.2]$$

donde t representa a un año entre 2015 y 2019.

Para los periodos anteriores, en los que no se dispone dato de MITMA, se utiliza un promedio entre el valor original de mercado y el precio estimado por el modelo de regresión de la Tabla 6.1. Calculado según la expresión:

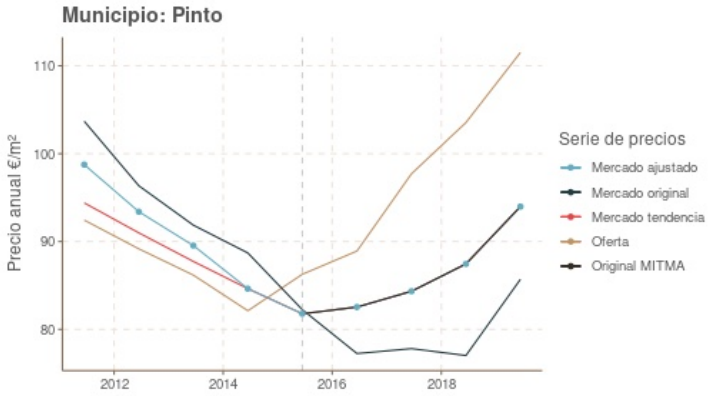
$$\lambda^M_{t,e} = \frac{P_{t,e}}{\omega \cdot \hat{P}^V_{t,e} + (1-\omega) \cdot (P_{t,e} \times \lambda_{2015,e})}, \forall t \in [2011..2014] \quad [6.3]$$

[6] Los ratios de ajuste del muestreo se han denominado factores de corrección dinámicos o λ.
[7] Las series de datos MITMA solo están disponibles a partir de 2015.

donde $\hat{P}^V_{t,e}$ es el precio del modelo de mercado original, y $P_{t,e} \times \lambda^M_{2015,e}$ el precio MITMA estimado a pasado, entre 2011 y 2015. El factor ω indica la proporción de la tendencia que se toma del modelo original, que en este caso se utiliza un valor de $\omega = 0,5$, y que está sujeto a revisión en futuras actualizaciones de la metodología.

El modelo final corrige el sesgo para cada macro-estrato zonal, definido como tipo (unifamiliar, plurifamiliar y zona). Por tanto, la serie final con el ajuste de sesgo se obtiene multiplicando el precio individual por la ratio de corrección λ del estrato.

Figura 6.6. Series de precios de mercado ajustada



Fuente: elaboración propia.

Como se observa en la Figura 6.6, los precios anteriores al 2015 se estiman a partir de los precios del modelo corregidos con la regresión de absorción de la oferta, y los posteriores, son ajustados mediante el dato MITMA de la zona. Para el caso de Pinto[8], en la tipología plurifamiliar, se aprecia la corrección de la tendencia en la serie de precios del mercado sobre el periodo 2016-2018.

Figura 6.7. Niveles de precios de series de mercado: EPF y MITMA

Fuente: elaboración propia.

[8]Se toma este ejemplo porque la serie de alquiler muestra un comportamiento incoherente con MITMA y oferta para el periodo entre 2016 y 2019.

Sin embargo, el precio de las rentas de las series de MITMA no se utiliza la misma magnitud que en la encuesta de la EPF, la cual contiene alquileres no sujetos a las declaraciones del IRPF, como son alquileres sociales o imputados. Se puede comprobar de manera gráfica sobre la Figura 6.7, la diferencia de escala entre las series de precios para el barrio Ciudad Jardín en Madrid[9].

Para efectuar el ajuste final a los niveles de precios de alquiler para la EPF, se calcula una tasa de ajuste final $\lambda_{t,e}$ a través de unos ponderadores w_e para cada estrato e, calculados como la proporción histórica entre el nivel estimado de precios MITMA y el precio de las series originales basadas en la EPF. El ponderador realiza un re-escalado de las series de precios de mercado, para que se ajusten a las medias de las series de la EPF desglosadas funcionalmente, y se elimine el sesgo de escala. El cálculo se realiza, por tanto, mediante la expresión siguiente:

$$\lambda_{t,e} = w_e \times \lambda_{t,e}^{M} \qquad [6.4]$$

Y por consiguiente, el precio final ajustado $\hat{P}_{t,e}$ se define como:

$$\hat{P}_{t,e} = \lambda_{t,e} \times P_{t,e} \qquad [6.5]$$

De forma alternativa, la expresión anterior se podría haber expresado en función del producto entre el factor MITMA (λ^M) y por la proporción de los pesos de la EPF (w_e), es decir:

$$\hat{P}_{t,e} = w_e \times \lambda_{t,e}^{M} \times P_{t,e} \qquad [6.6]$$

El análisis de los valores λ^M tiene una utilidad adicional, permite medir el grado de sesgo incurrido en el modelo de precios original.

[9] Se toma una zona urbana con buen nivel de muestra, en la que se aprecia notablemente la diferencia de escala entre el precio MITMA y el de alquiler.

6.3 Resultados

Para evaluar los objetivos perseguidos de incorporar frecuencia mensual a las series, y corregir la coherencia geográfica de los precios de mercado, se han analizado los resultados del modelo final en torno a cuatro cuestiones:

- El método es capaz de reducir el sesgo zonal del modelo de mercado original.
- Dado que los modelos de *bagging* son propensos a los sesgos, se evalúa su presencia en los precios finales de oferta y de mercado.
- Control la coherencia mensual y anual de las series de precios.
- Confirmar la capacidad de generalización en el tiempo del modelo de mercado.

6.3.1 Eficacia en el control del sesgo zonal

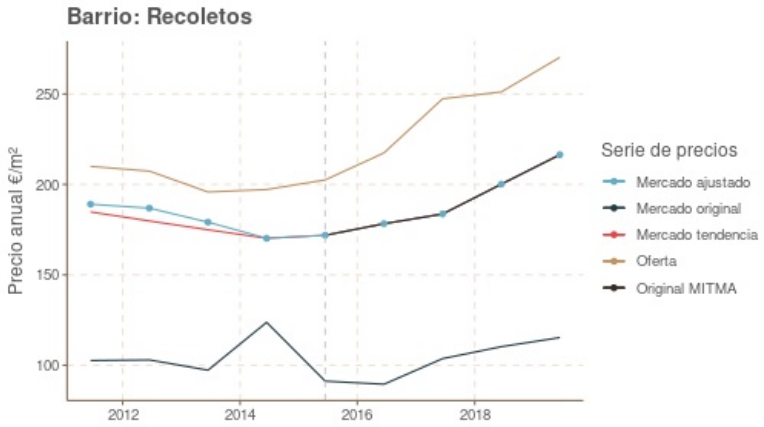
En esta sección se abordarán los dos sesgos principales del modelo, y que son la fuente principal de las distorsiones en los precios del modelo de alquiler:

- Sesgos debidos a la composición, y que se producen por que la calibración entre las poblaciones de oferta y alquiler no tiene encuentra los niveles de precios de las zonas.
- Sesgos propios de los precios de la oferta, que se manifiestan principalmente en zonas con menor nivel de muestra, donde hay una gran variabilidad del precio por unidad de superficie o la estimación está más expuesta a variables omitidas.

Para ilustrar de forma gráfica el método de ajuste, se representa a modo de ejemplo el resultado para del barrio de Recoletos en Madrid en la Figura 6.8 [10] (para ver todos los ejemplos, consúltese https://github.com/davidreyblanco/idx/tree/master/hedonic-final/unbias-fixed y los enlaces del Anexo I). Se observa un descenso anormal en el precio de mercado para el año 2015, sin una razón de mercado plausible que pueda justificarlo, y que es incoherente con los precios medianos del MITMA. La serie del modelo “final” propuesto, representada en la imagen como “Mercado ajustado”, corrige eficazmente la tendencia y ofrece una serie de precios de mercado coherente con el resto de series originales (MITMA y oferta).

[10] Se toma el caso de Recoletos para evidenciar el primero de los problemas (composición zonal), ya que cuenta con una muestra de anuncios numerosa y muestra una serie de precios de alquiler incoherente con los precios de oferta y los registrados en MITMA, particularmente para el año 2014.

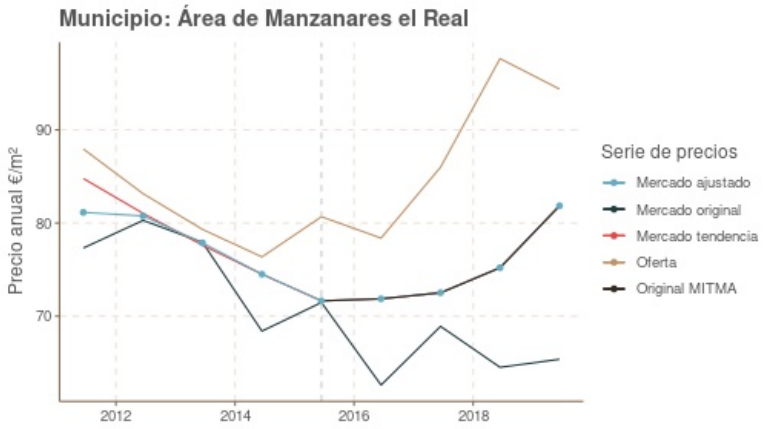
Figura 6.8. Series de precios vivienda plurifamiliar: barrio de Recoletos



Fuente: elaboración propia.

Es habitual que las zonas con muestras más pequeñas e irregulares den lugar a series de precios con variabilida espuria como podemos ver en Kokot y Bas (2015) o Eurostat (2013). Este fenómeno también se aprecia en nuestro caso, por ejemplo, con los datos para el municipio de la sierra madrileña de Manzanares el Real (municipio rural con un mercado inmobiliario poco activo), mostrados en la Figura 6.9. Se observa como la serie corregida controla las anomalías de modelo de mercado hasta 2017, y la excesiva caída de precios medios en oferta de 2019.

Figura 6.9. Series de precios vivienda plurifamiliar: Manzanares el Real

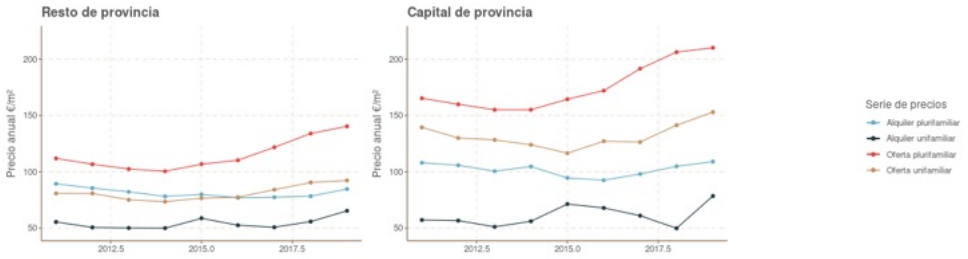


Fuente: elaboración propia.

De la misma manera, se corrige la incidencia del efecto de composición. Las Figuras 6.10 y 6.11 muestran los precios agregados originales y corregidos, en las que se ajustan eficazmente los efectos mencionados en los párrafos anteriores. Debido a su mayor representatividad, la corrección ligeramente mejor para los

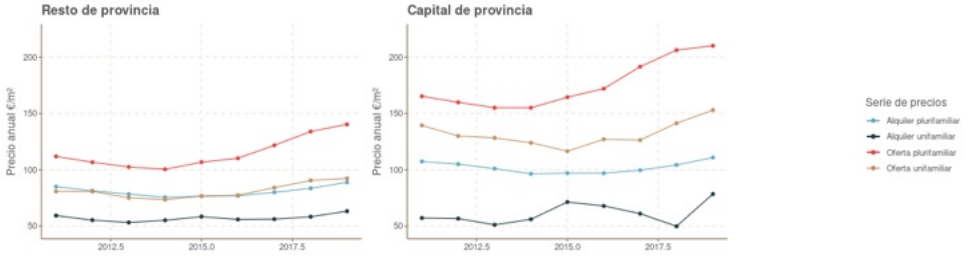
precios para viviendas plurifamiliares que en las unifamiliares.

Figura 6.10. Series de precios agregados, originales



Fuente: elaboración propia.

Figura 6.11. Series de precios agregados, corregidas



Fuente: elaboración propia.

Desde un punto de vista numérico, la Tabla 6.2 muestra como los precios de mercado ajustados por el modelo reducen de forma significativamente la variabilidad de los precios originales. Se aprecia, además, que se preservan los parámetros principales como la media y cortes cuantílicos.

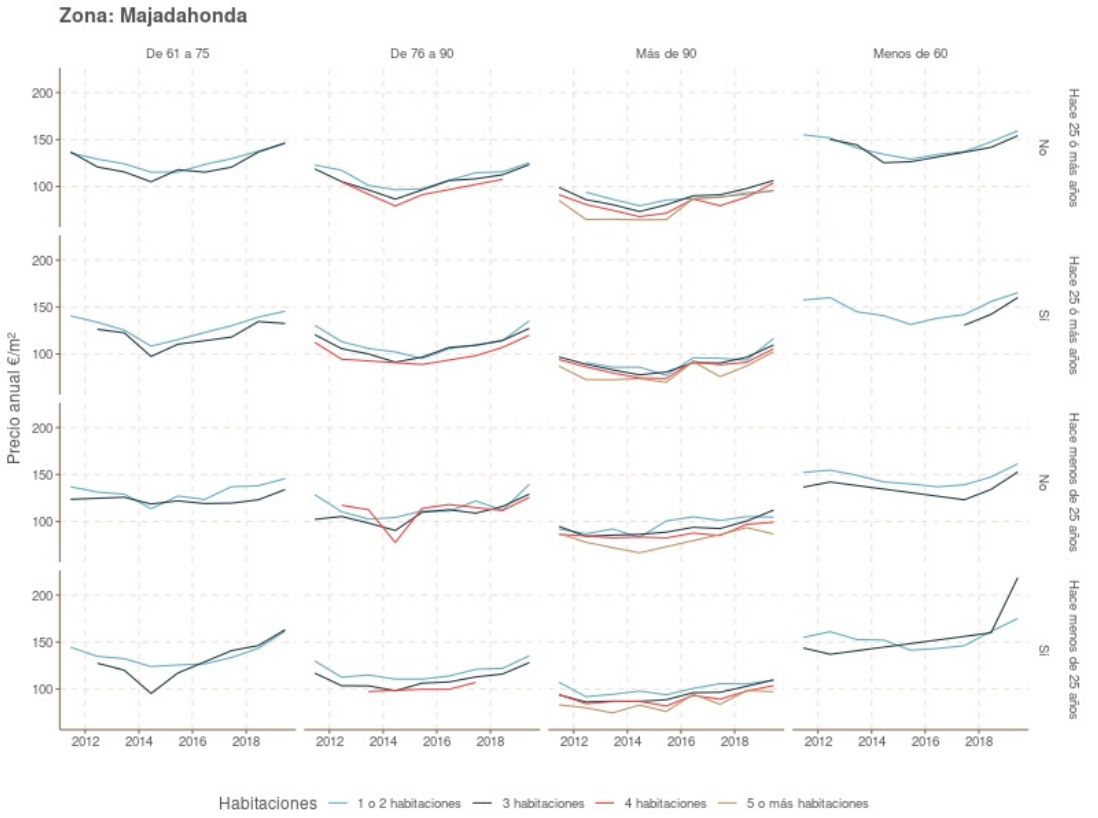
Tabla 6.2. Comparativa de agregados de series antes y después del ajuste zonal

Capital	Tipo	Paso	Media	Desv	Min	p25	p50	p75	Máx.
No	Plurifamiliar	Ajustado	80.73	4.44	75.59	76.91	79.97	83.54	88.94
		Original	81.43	4.33	76.97	78.16	79.96	84.72	89.41
	Unifamiliar	Ajustado	57.27	2.95	53.27	55.34	56.23	58.41	63.27
		Original	54.40	5.13	49.99	50.67	52.60	55.84	65.40
Sí	Plurifamiliar	Ajustado	102.08	5.13	96.46	97.01	100.99	105.03	110.90
		Original	102.02	5.90	92.58	98.03	104.70	105.84	109.04
	Unifamiliar	Ajustado	61.12	9.62	49.88	56.12	57.27	67.94	78.54
		Original	61.12	9.62	49.88	56.12	57.27	67.94	78.54

Se verifica gráficamente la preservación de la coherencia al desagregar las series según criterios funcionales. A modo de ejemplo se toma el municipio

de Majadahonda, representativo de las zonas residenciales del extrarradio de Madrid, cuyos precios originales (Figura 6.12) muestran una coherencia temporal consistente con las series corregidas de la Figura 6.11. Se observa cierto grado de cointegración entre series y una mayor variabilidad en los segmentos de mayor superficie, como aquellos con 4 y 5 o más habitaciones, debido a una menor representatividad estadística en estos segmentos.

Figura 6.12. Precios de mercado ajustados por número de habitaciones, superficie, antigüedad y si dispone de piscinas



Fuente: elaboración propia.

Finalmente, se estudia de forma cuantitativa el sesgo del resultado del ajuste zonal, para lo que se definen las métricas de $Bias_{zonal,s}$ y $Bias_{funcional,e}$, que representan el nivel de divergencia de los precios.

El primero, referido al sesgo zonal de cada zona s, y calculado como el nivel de discrepancia de las variaciones de la serie de mercado con respecto a la serie MITMA, pretende medir el nivel de diferencia en la tendencia de ambas series para las distintos estratos zonales definidos por: zona, tipo de vivienda y año. El cálculo de discrepancia se realiza según la siguiente expresión:

$$Bias_{zonal,s} = \frac{\sum_{t=2019}^{2015} w_{t,s} \cdot \left(\Delta \hat{P}_{t,t-1,s} - \Delta P^{M}_{t,t-1,s}\right)}{\sum_{t=2019}^{2015} w_s} \text{ , } \forall t \in [2015..2019], s \in S \quad [6.7]$$

donde $\Delta \hat{P}_{t,t-1,s}$ representa la variación logarítmica anual del precio de mercado para la zona s en un año t, $\Delta P^{M}_{t,t-1,s}$ es la correspondencia variación logarítmica de los precios medianos de MITMA y $w_{t,s}$ el peso poblacional de este estrato zonal-temporal.

El sesgo funcional $Bias_{funcional,e}$, representa las diferencias de precios medios ponderados de los datos generados por el modelo ($\overline{P_e}$) con respecto a los precios medios ponderados del resultado del ajuste zonal ($\overline{\hat{P}_e}$). El estrato funcional e se corresponde con la parte de la población de la celda de unidad mínima utilizada en la calibración, es decir, la combinación de las siguientes variables[11]: *CAPROV, TIPOCASA, NHABIT, TAMAMU, ANNOCON, TIPOEDIF, ZONARES, DENSI, HASBOXROOM, HASPARKINGSPACE, HASSWIMMINGPOOL, HASAIRCONDITIONING, SUTC y SUPERF*. La fórmula de cálculo de la métrica de sesgo funcional es la siguiente:

$$Bias_{funcional,e} = \overline{\hat{P}_e} - \overline{P_e} \quad [6.8]$$

La Tabla 6.4 y la Tabla 6.3 muestran los descriptivos de las medidas de sesgo zonal y funcional respectivamente. Se observa que el sesgo funcional medio es prácticamente cero para viviendas plurifamiliares, las cuales representan más del 95% de la población. Solo en el caso de la vivienda unifamiliar fuera de la capital los valores del nuevo modelo son ligeramente superiores (2,66), y en el caso de Madrid es 0,0 porque no se realiza un ajuste zonal. Por tanto, se confirma la eficacia del método propuesto para eliminar el sesgo zonal, al ser prácticamente cero y con un nulo nivel de variabilidad.

[11] Véase el epígrafe 3.2.2.1, en el Capítulo 3, para más información sobre el significado y mayor información de las variables.

Tabla 6.3. Sesgo funcional después de ajuste funcional

Es capital	Tipo	N. estratos	Peso	Bias	Varianza bias
No	Plurifamiliar	90.705	43,1%	-0.27	90.16
	Unifamiliar	27.713	4,1%	2.66	299.67
Sí	Plurifamiliar	130.693	52,4%	0.00	175.36
	Unifamiliar	7.239	0,4%	0.00	0.00

Fuente: elaboración propia

Tabla 6.4. Sesgo zonal después de ajuste zonal

Es capital	Tipo	N. estratos	Peso	Bias	Varianza bias
No	Plurifamiliar	1.300	46,3%	0.01	0
	Unifamiliar	1.300	4,7%	0.01	0
Sí	Plurifamiliar	2.240	48,7%	0.01	0
	Unifamiliar	1.635	0,3%	0.00	0

Fuente: elaboración propia

6.3.2 Sesgos en los precios finales de oferta y mercado

Existen dos posibles fuentes de sesgo en los precios de oferta:

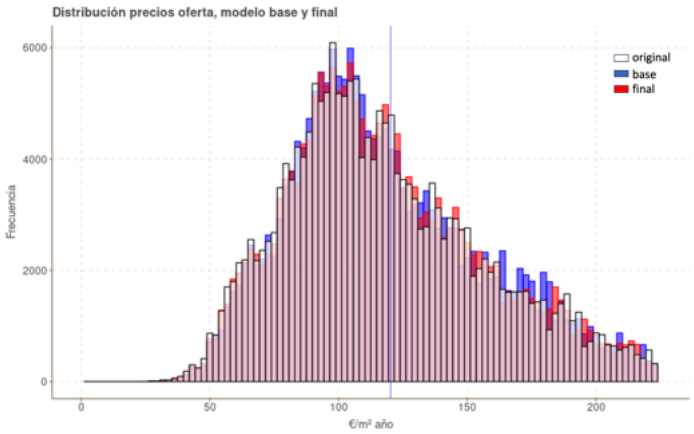
- Las diferencias en la distribución de valores de mercado y de oferta (Kolbe *et al.*, 2021; Ohnishi *et al.*, 2011; Shimizu *et al.*, 2016).
- Los estructurales de los modelos de tipo *Random Forests* (Hastie *et al.*, 2017), propensos a excluir los valores más extremos (Antipov y Pokryshevskaya, 2012), cuestión descrita en profundidad en el Anexo 6a del presente capítulo.

Al comparar los histogramas de frecuencias de precios ponderados[12] (Figura 6.13) no se observan diferencias significativas entre las tres valores: los originales (los presentes en los registros de originales de idealista), los del modelo de oferta del modelo de correspondencia (denominados “base” en la gráfica) y los finales (calculados por el modelo final de oferta). No obstante, se aprecia ligeramente una mayor concentración de los valores de los modelos alrededor del centro de masas, que coincide la propuesta de Antipov y Pokryshevskaya (2012), sobre que estos métodos tienden a eliminar los valores más extremos en los modelos, y por tanto a ofrecer peores estimaciones en estos casos[13] .

[12] Los valores de la representación se encuentran ponderados según sus pesos muestrales.

[13] Lo que no significa que los modelos funcionen incorrectamente, sino que el modelo no considera que son observaciones asociadas a comportamientos generalizables.

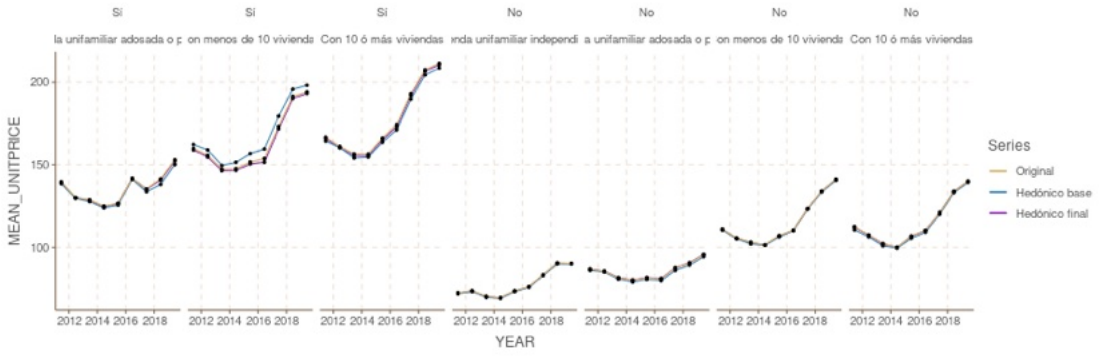
Figura 6.13. Distribución de precios de oferta



Fuente: elaboración propia.

Cuando se ponen en común las series temporales de precios de oferta (Figura 6.14), para los precios desglosados por capital de provincia y por tipo de edificio, se observa que los modelos base y final mantienen la tendencia de la serie de precios de oferta original. Los valores de las series tampoco muestran discrepancias en valores, exceptuando el caso de Madrid para edificios con menos de 10 viviendas, en el que el modelo final ajusta el sesgo del modelo base.

Figura 6.14. Precios de oferta desglosado por capital de provincia y tipo de edificio



Fuente: elaboración propia.

Para realizar el análisis en profundidad, y como no se dispone una correspondencia a nivel de registros observados y estimados[14], el sesgo se estima como el agregado de las contribuciones ponderadas de los sesgos de los estratos. En este caso, se estima de la forma siguiente:

[14] El valor observado de alquiler correcto sería el precio por el que se ha alquilado el inmueble en oferta.

$$Bias_Y = \sum_{e \in E} w_e \cdot Bias_Y(e) \quad [6.9]$$

donde $Bias_Y(e)$ es el sesgo del precio de oferta para el estrato funcional e[15], ponderado según el factor de elevación poblacional w_e del estrato.

Las medidas de sesgo de ambos modelos (Tabla 6.5) indican que tienden a infravalorar ligeramente, aunque el modelo hedónico final reduce a una tercera parte el sesgo del modelo original. Lo que indica que el modelo final es especialmente restrictivo con los valores extremos superiores.

El fenómeno de sesgo negativo en oferta es también identificado por Kolbe (2021), que lo atribuye a que los precios implícitos[16] son más dominantes en los modelos de oferta que en los modelos sobre operaciones reales.

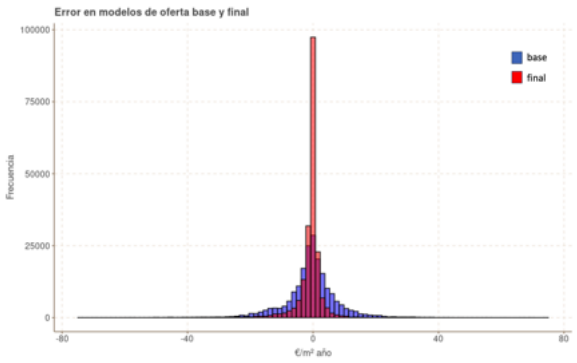
Tabla 6.5. Sesgo modelos hedónicos de oferta

Model	Media	Mediana	Pct. Media	Pct. Mediana	Desviación
Hedónico base	-0.97	-1.44	-0,67%	-0,99%	7.94
Hedónico final	0.57	0.14	0,39%	0,10%	3.36

Fuente: elaboración propia

De forma gráfica, se confirma el efecto de reducción de los errores en el modelo de oferta final 6.15. E indica cómo el uso de más variables afecta la variabilidad de los errores[17], a pesar de que el sesgo de los errores de ambos modelos sea muy cercano a cero.

Figura 6.15. Distribución de sesgos en modelos de oferta



Fuente: elaboración propia.

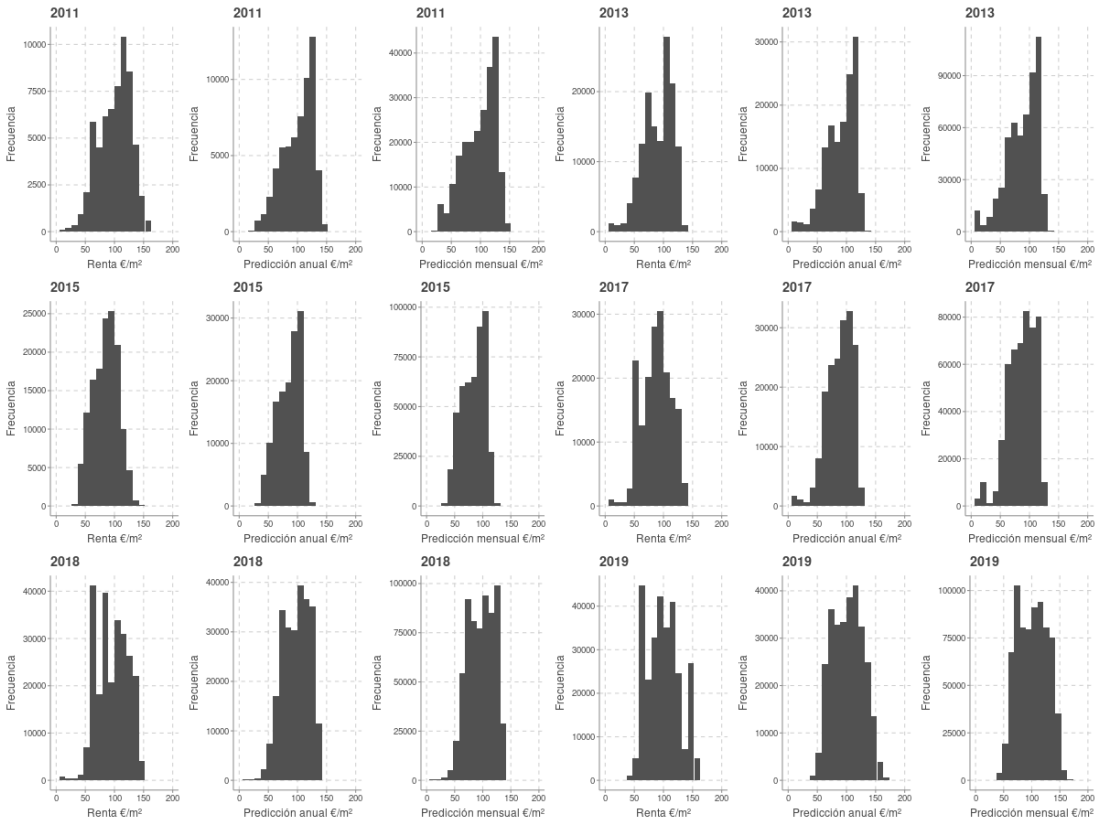
[15] Se sigue la misma estratificación funcional del apartado 6.3.1

[16] Los precios implícitos se refieren al valor económico de los factores no incluidos en el modelo, como son estado de conservación o cualquier otra variable omitida.

[17] Lógicamente, cuantas más variables estén disponibles menor será el riesgo de incurrir en sesgos de omisión de variables.

Para el caso de los precios de mercado, en la Figura 6.16 se comparan los histogramas de frecuencias para los precios de la EPF y los modelos, usando los elevadores muestrales de la calibración. Se observa que mientras que los precios de los modelos anuales y mensuales son equivalentes, los precios de la EPF son ligeramente diferentes. Por otra parte, la distribución de valores en cada conjunto varía en el tiempo, con un histograma más equilibrado en los últimos años de la serie, que se puede relacionar con un mayor tamaño muestral.

Figura 6.16. Distribuciones estimaciones del precio del alquiler



Fuente: elaboración propia.

Para estudiar en detalle los valores numéricos asociados a la figura anterior, se han representado los parámetros principales del precio ponderado de alquiler en la Tabla 6.6. No se aprecian diferencias significativas, ni en los cortes por cuantiles ni en las medias, aunque los precios del modelo hedónico son ligeramente superiores a los originales. Por otra parte, tal y como se observaba en los precios de oferta, la mayor desviación típica del modelo final indica que se reduce, de forma general, la variabilidad final.

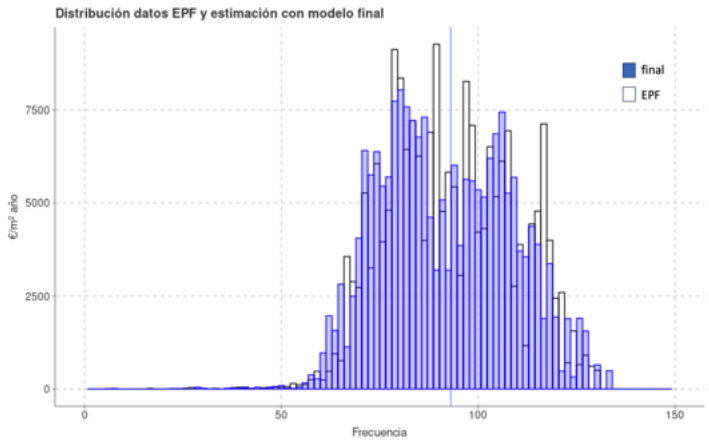
Tabla 6.6. Parámetros precios de alquiler anuales originales)

	EPF					Modelo Final				
Año	**Media**	**Desv**	**P25**	**P50**	**P75**	**Media**	**Desv**	**P25**	**P50**	**P75**
2011	99.61	571.13	81.15	100.77	117.81	99.74	507.76	83.39	102.27	118.74
2012	96.15	608.17	76.15	98.55	112.37	96.60	562.39	78.57	97.90	114.41
2013	91.27	463.69	74.49	93.33	108.16	91.54	433.12	76.14	93.61	108.87
2014	89.96	721.26	72.12	86.79	103.83	91.22	701.70	72.64	89.40	106.80
2015	85.78	388.83	71.91	86.55	100.24	86.11	336.69	72.80	88.38	101.09
2016	82.73	360.90	70.58	81.81	97.71	83.33	321.75	71.56	85.35	96.94
2017	86.18	421.04	73.64	85.35	99.53	86.54	373.09	73.20	88.34	100.89
2018	89.77	536.61	72.00	88.63	104.67	90.25	509.73	74.07	90.26	108.00
2019	94.65	546.02	75.82	95.79	109.48	95.63	535.06	78.21	95.63	111.92

Fuente: elaboración propia

De forma gráfica, la distribución de los valores de los distintos casos[18] se muestran en la Figura 6.17. Se observa que los valores estimados mantienen una distribución similar a los valores observados en la EPF, existiendo un muy ligero sesgo a infravalorar por parte del modelo. Existe una mayor irregularidad en la distribución, si se compara con la gráfica de oferta de la Figura 6.17, probablemente porque la fuente destino sea más irregular, similar a lo indica Kokot (2015) cuando compara la estabilidad de series de alquiler de precios de oferta y registros oficiales.

Figura 6.17. Distribución de precios de alquiler

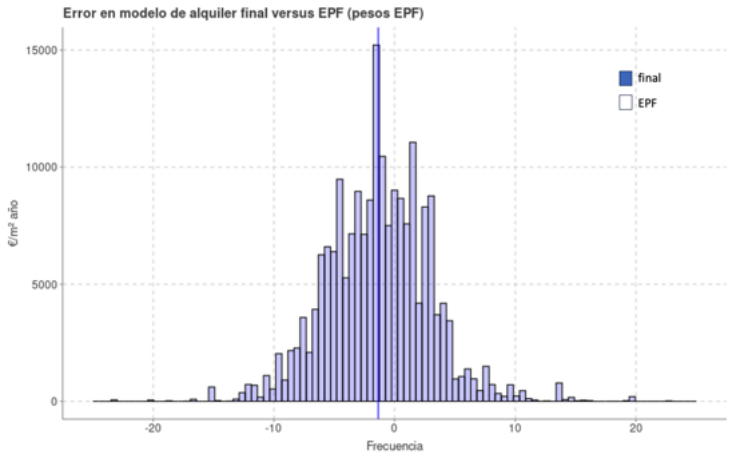


Fuente: elaboración propia.

[18]Para este caso densidad de precios ponderadas usando los pesos de la EPF, para estimar las funciones de distribución empíricas original y del modelo, en enfoque se basa una aproximación numérica inspirada en (Monahan, 2011).

Los errores del modelo se concentran en torno a cero, como se ve en la Figura 6.18, aunque al contrario del modelo de oferta, existe una menor concentración de errores nulos.

Figura 6.18. Distribución de errores del modelo de alquiler final



Fuente: elaboración propia.

Los valores de sesgo de los errores, mostrados en la Tabla 6.7, indican que el modelo final tiene un sesgo negativo (es decir que el modelo tiende a infravalorar). La desviación típica un 60% mayor que para el caso de la oferta, confirma la mayor dispersión de errores de la Figura 6.18.

Tabla 6.7. Sesgo modelos hedónicos de alquiler respecto a la EPF

Model	Media	Mediana	Pct. Media	Pct. Mediana	Desviación
Hedónico final	-1.53	-1.48	-1,542%	-1,494%	5.39

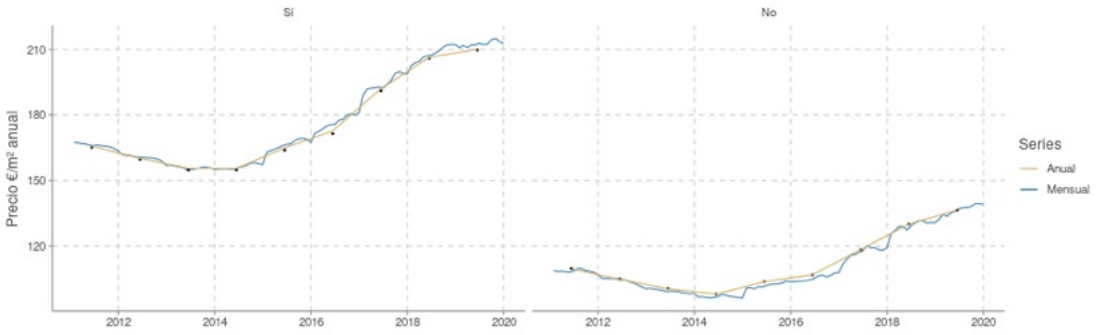
Fuente: elaboración propia

6.3.3 Coherencia temporal entre series anuales y mensuales

Puesto que las series mensuales finales de mercado se estiman con el modelo de mercado que se calcula sobre datos anuales, es necesario comprobar que esta diferencia de escalas temporales afecten a las series generadas. Cuando se concilian series con distintas frecuencias y múltiples periodos es habitual la presencia de discontinuidades (Hood, 2005), Chen y Andrews (2008) recogen que los cambios entre noviembre y febrero suelen ser particularmente superiores a los del resto del año .

Se puede comprobar gráficamente, en la Figura 6.19, que la variación entre los meses diciembre y enero es efectivamente superior a la observada en el resto del año.

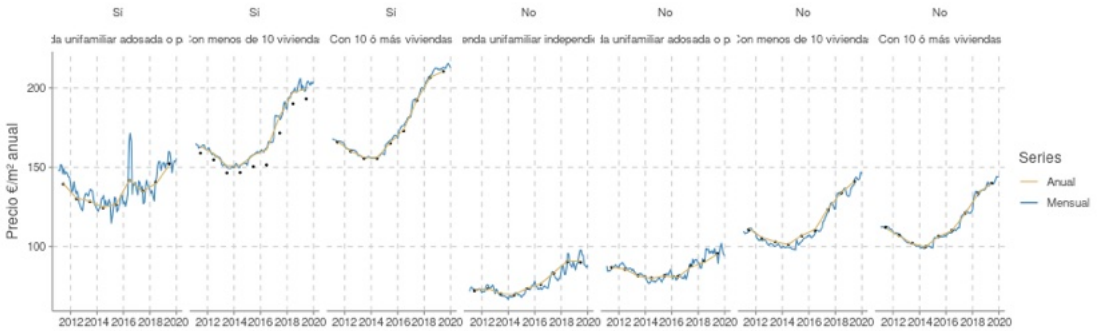
Figura 6.19. Precios de mercado precios mensuales y anuales, desglose capital y provincia



Fuente: elaboración propia.

El patrón anterior se acentúa a medida que se trabaja con estratos más pequeños, como muestra el desglose de precios medios por capital/provincia y tipo de edificio en la Figura 6.20. A este respecto, se puede establecer una relación con los resultados de los modelos de mercado, para los que cuanto menor es el tamaño del estrato más frecuente es la irregularidad de los resultados.

Figura 6.20. Precios de mercado precios mensuales y anuales, desglose tipo de edificio y capital/provincia



Fuente: elaboración propia.

Se puede confirmar el mismo comportamiento que con las series más agregadas, los cambios más abrupto en las series se corresponden, exactamente, a los periodos entre octubre y febrero, tal y como muestran los valores de la Tabla 6.8.

Tabla 6.8. Variación precio del alquiler con su mes anterior

Mes	Variación	Mes	Variación
Enero	3,32%	Julio	0,33%
Febrero	0,57%	Agosto	0,28%
Noviembre	0,52%	Septiembre	0,22%
Diciembre	0,42%	Abril	0,20%
Octubre	0,40%	Junio	0,20%
Marzo	0,39%	Mayo	0,18%

Fuente: elaboración propia

Estas discontinuidades no parecen atribuibles al proceso de reponderación, ya que, como se muestra la Figura 6.21, se observa de forma recurrente un cambio de escala de precios entre noviembre y enero del año posterior, para los pesos originales de la EPF y los calculados por la reponderación. Por otra parte las series mensuales y anuales son muy similares, con diferencias entre las medias mensuales y los valores anuales, que oscilan entre el 0,47% y el 1,48%.

Figura 6.21. Precios de mercado precios mensuales y anuales, viviendas plurifamiliares y pesos EPF

Fuente: elaboración propia.

Se puede concluir que los datos mensuales de mercado producidos por el modelo final reproducen las series anuales, aunque no exactamente, y tampoco capturan los cambios intermensuales, especialmente entre el fin y principio de año. Por tanto, como propone Eurostat (2015), es necesario un proceso de conciliación de series anuales y mensuales, para mitigar estos efectos indeseados.

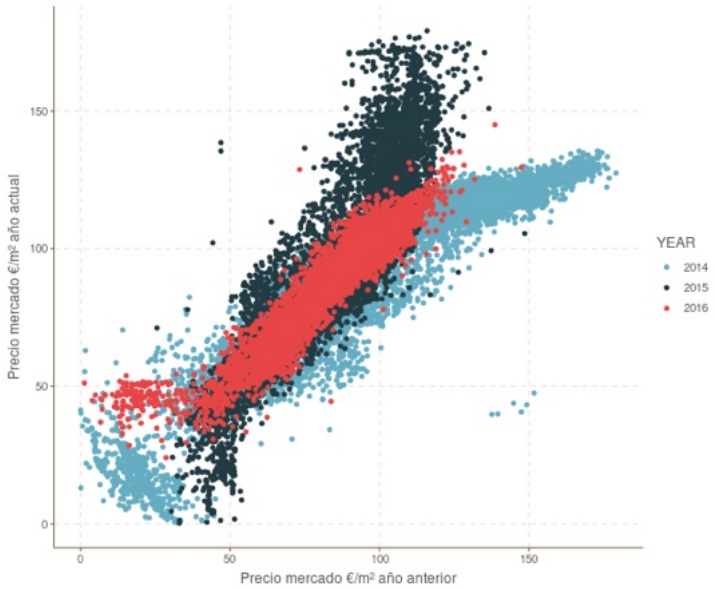
6.3.4 Capacidad de generalización para periodos futuros

Tal y como argumentan Shimizu *et al.* (2016), los modelos basados en datos de oferta cuentan con la ventaja de la inmediatez de la información. Sin embargo, aunque el dato del portal de internet está disponible de forma inmediata, el dato de mercado no, y se actualiza por el INE de forma anual y con un retraso de publicación de casi un año más. Esto resta, de forma estricta, aplicabilidad al modelo, al no permitir estimaciones de precio de mercado actualizadas.

Este problema se reduciría si la función que relaciona los precios del portal y los de alquiler fuera persistente en el tiempo[19], y por tanto, el modelo de un año sería válido para proyectar los datos de alquiler del año siguiente.

Para comprobar la hipótesis anterior, se estiman los precios de mercado para los años 2014, 2015 y 2016 usando el modelo de conversión correspondiente al periodo del año anterior. La Figura 6.22 demuestra gráficamente que existe una correlación entre ambos valores, en ella se relacionan el precio para un inmueble usando los modelos de años consecutivos. La pendiente variable según el año indica que relación varía ligeramente en el tiempo, pero sigue siendo fuerte.

Figura 6.22. Relación precios de mercado por el modelo de su periodo y el modelo del año anterior



Fuente: elaboración propia.

La muestra conformada por 27.766 observaciones ofrece un coeficiente de correlación de Pearson, entre años consecutivos, del 0,851, confirmándose la hipótesis de una fuerte relación persistente en el tiempo. Esta correlación es, en

[19] En este caso se mide la persistencia solamente entre años contiguos.

realidad, mucho más fuerte si observamos el desglose para cada año, como se aprecia en la Tabla 6.9, variando entre 0,86 y 0,94.

Tabla 6.9. Coeficiente de correlación de Pearson precio de oferta con modelos de años consecutivos

Año	Coef. correlación de Pearson
2014	0.91
2015	0.86
2016	0.94

Fuente: elaboración propia

Para confirmar que la relación entre magnitudes es significativa, se desarrolla un modelo lineal de mínimos cuadrados ordinarios que estima el precio de alquiler del año en curso, según el precio estimado del año anterior, y definido como:

$$\hat{P}_e(t) = \beta_0 + \beta_1 P_e(t-1) + \sum_{a=1}^{N} \gamma_a D_a + \varepsilon_t \qquad [6.10]$$

donde $\hat{P}_e(t)$ representa el precio de alquiler estimado para un estrato e y un año t, D_a es una variable *dummy* que toma valor 1 si el año t se corresponde a a, y cuya función es controlar las variaciones específicas de cada año. Finalmente, ε_t se refiere al término de error aleatorio.

La Tabla 6.9 muestra los coeficientes del modelo, donde *ANNUAL_PR_RENT_T1* se corresponde a β_1, mientras que *YEAR2015* y *YEAR2016* se refiere a los coeficientes γ_a de las *dummy* de tiempo D_a en la expresión [6.10]. Se observa que los *p-valores* de todos los coeficientes de las variables independientes (Tabla 6.10) son significativos, con un grado de significación inferior al 0,001. Lo cual confirma que las variables independientes estudiadas aportan información relevante y valiosa para explicar y predecir el comportamiento de la variable dependiente.

Tabla 6.10. Modelo de relación precios estimado con modelo del año anterior

	Estimación	**std.error**	**t value**	**Pr(>|t|)**	**signif.**
(Intercept)	9.50	0.31	31	< 2e-16	***
ANNUAL_PR_RENT_T1	0.88	0.00	286	< 2e-16	***
YEAR2015	9.86	0.19	52	< 2e-16	***
YEAR2016	4.99	0.22	23	< 2e-16	***

Fuente: elaboración propia

Códigos signif.: *** 0,001 ** 0,01 * 0,05 . 0,1 1
Error estándar de los residuos: 14 sobre 27766 grados de libertad (DF)
R^2: 0,74893, R^2 ajustado: 0,74891
Estadístico-F: 27605 sobre 3 y 27762 DF, p-value: < 2,2e-16
Num. observaciones: 27766

En este capítulo, se ha presentado un método que corrige eficazmente el sesgo producido por la falta de información zonal del modelo de mercado. Los precios ajustados se utilizarán para el cálculo de las series de precios de los distintos estratos de la población, que posteriormente se aplicarán en la construcción del índice de precios final, presentado en el Capítulo 8. Sin embargo, los modelos calculados hasta este momento ofrecen datos anuales, por lo que se desarrollará un método para desagregar los datos a frecuencia mensual, y que se presentará en el Capítulo 7.

Anexo 6a. Identificación y corrección de sesgos

Las estimaciones de un modelo de aprendizaje automático de árboles tiende a ser insesgado en el sentido de que la suma de los errores (observados contra los estimados) es cercano a cero (Hastie *et al.*, 2017). Sin embargo, los modelos de regresión calculados con estos métodos pueden arrojar resultados sesgados en un sentido diferente (Zhang y Lu, 2012): los valores pequeños se sobreestiman y los valores altos se infraestiman. Para muchos propósitos es importante cualificar correctamente los casos extremos de la distribución. En el caso de las valoraciones inmobiliarias existe una gran diversidad de mercados, y dentro de un submercado existen inmuebles singulares, por tanto este tipo de modelos deben ser capaces de tratar los casos más extremos.

En un modelo de regresión el error del modelo es la suma de varianza aleatoria (ruido blanco o ε) del sesgo del predictor y la varianza del predictor, donde los dos últimos componentes se denominan riesgo de la función de regresión, descrito en [6.11]. Breiman (1996) demuestra que la técnica de *bagging* puede reducir de forma eficaz la varianza del predictor, pero no actúa sobre el sesgo, siendo este último el factor dominante del riesgo[20] de los modelos. Por otra parte, la reducción de la varianza da lugar a que este tipo de métodos no traten correctamente los valores extremos y por tanto pueda ser necesaria una corrección de sesgo.

$$R[\hat{f}(x)] = Bias[\hat{f}(x)] + \sigma^2[\hat{f}(x)] \qquad [6.11]$$

donde $R[\hat{f}(x)]$ se refiere al riesgo del estimador $\hat{f}(x)$, $Bias[\hat{f}(x)]$ es su sesgo, y $\sigma^2[\hat{f}(x)]$ la varianza del mismo.

Existe una segunda fuente de sesgo de los modelos estadísticos producida por las transformaciones de la variable dependiente. En nuestro caso, se transforma la variable de precio a escala logarítmica, lo que es útil en los modelos lineales para reducir la heterocedasticidad en la variable respuesta, pero puede dar lugar a sesgo en la magnitud estimada al revertir la transformación de la variable dependiente.

Se puede establecer una medida numérica de sesgo como la diferencia entre los valores observados, $\hat{f}(observado)$, y los valores estimados por el modelo, $\hat{f}(estimado)$. Por tanto, se define el error de un estimador $\epsilon[\hat{f}(x)]$ como:

[20] El riesgo mide los dos aspectos clave en el ajuste de los modelos: la varianza y el sesgo.

$$\epsilon[\hat{f}(x)] = \hat{f}(observado) - \hat{f}(estimado) \quad [6.12]$$

Y el sesgo como esperanza de los errores del modelo en:

$$Bias[\hat{f}(x)] = E\{\epsilon[\hat{f}(x)]\} \quad [6.13]$$

En general, el control de sesgo se afronta desde dos perspectivas, el primero basado en la escala del punto (*point-scale*) o basado en la escala de la distribución (*distribution-scale*). El primer enfoque ajusta los valores estimados para que la desviación entre la estimación y el valor observado sea mínimo, y el segundo ajusta la distribución de la variable estimada de forma que las distribuciones acumuladas, del modelo y la variable dependiente, se correspondan. Las correcciones de sesgo sobre los valores puntuales se han tratado de forma extensa en la literatura estadística (Giffen *et al.*, 2022; Hort *et al.*, 2022; Pagano *et al.*, 2022)

Para *Random Forests* el modo más sencillo de controlar el sesgo es a través los registros OOB, descritos en detalle en el Anexo II del Capítulo 3. Por ejemplo, Liaw y Wiener (2002) proponen usar una regresión lineal sobre este conjunto datos para realizar el ajuste, o Zhang y Lu (2012) propone usar una regresión generalizada con la muestra completa. Existen otros método alternativos que usan un segundo modelo de *Random Forests* para controlar el efecto del sesgo, denominados *"one-step boosted forests"* (OSFB) (Zhang y Lu, 2012), que en general muestran mejores resultados que los método basados en los registros OOB.

Los método basados en la corrección de la distribución parten de encontrar un modelo de correspondencia entre las dos distribuciones, y no se han aplicado tan extensamente en todos los campos, siendo quizá la meteorología el único campo donde se ha aplicado de forma más frecuente.

Capítulo 7

Desagregación temporal

> *"El estudio de la economía normalmente nos muestra que el mejor momento para comprar fue el año pasado."*
>
> — Woody Allen

7.1 Introducción

El marco metodológico propuesto en esta Tesis parte de modelos hedónicos que generan precios con frecuencias anuales y mensuales. De la misma manera que en el Capítulo 6 se ha desarrollado un método que asegura la coherencia geográfica de los precios, en el presente capítulo se aplica un método que asegura la consistencia temporal entre las series de alta y baja frecuencia.

Trabajar con frecuencias distintas conlleva dos problemas: el primero, que los datos mensuales tienden a concentrarse en torno a la media anual, lo que provoca discontinuidades entre los valores de los meses finales de un año y los iniciales del siguiente; y el segundo, que las medias anuales no se corresponden exactamente con el promedio de los valores de la serie mensual. Estas cuestiones se resuelven integrando ambas series mediante la reconciliación de la serie anual y la mensual.

La cuestión plantea varios retos, como la selección de las series de referencia[1] para estimar el precio de una serie de alta frecuencia mensual a partir de una serie de baja frecuencia anual. Pero además, la identificación de información de alta frecuencia actualizada también es problemática, ya que los organismos públicos suelen ofrecen información del mercado con varios meses o años de retraso.

Como se describe más adelante en este capítulo, las series de referencia candidatas tienen retrasos temporales y manejan escalas distintas. Por lo tanto,

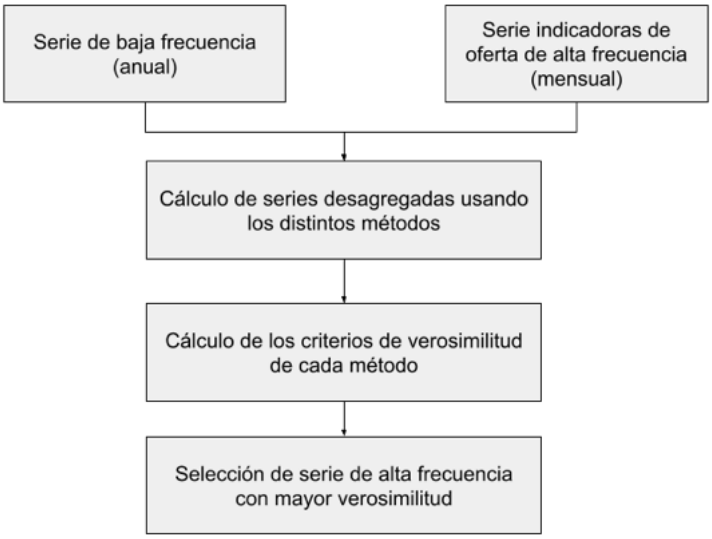
[1] Podrían ser varias series en el caso de la aproximación multivariante.

dada la diversidad en la naturaleza de las series de los estratos poblacionales, es necesario adaptar el método de desagregación en función de cada caso.

Dado que existen múltiples opciones para la reconciliación de los datos, la metodología propuesta toma las series de baja frecuencia y las desagrega utilizando varios métodos. Seleccionando aquellas que, según un criterio de verosimilitud, muestre el mejor ajuste al valor original y ofrezca las series con más alta calidad (reduciendo casos de discontinuidad o comportamientos erráticos de alta frecuencia).

El proceso general sobre la serie de tiempo de una zona se describe en la Figura 7.1. El método calcula las series desagregadas a través de distintas técnicas, partiendo del dato anual de mercado y las series mensuales de oferta. Sobre las series calculadas, se estima su verosimilitud en función de varios criterios de comprobación de calidad, por ejemplo, la variación interanual o las diferencias entre los valores de fin e inicio de año. Finalmente, se toma aquella cuyo valor de verosimilitud sea el más alto.

Figura 7.1. Algoritmo de cálculo de selección de series por máxima verosimilitud



Fuente: elaboración propia.

El criterio de selección del mejor método se basa en las condiciones de calidad de series temporales más comúnmente utilizadas por las agencias estadísticas. La decisión final sobre el método a utilizar se construye como una combinación de los criterios individuales, mediante un estimador de máxima verosimilitud (MLE) (Bickel y Doksum, 2015) que asegura el máximo cumplimiento de las condiciones de calidad individuales de las series. Los MLE son una herramienta indispensable en muchas técnicas de modelización estadística, especialmente ante datos que no siguen una distribución normal, y por tanto incumplen la restricción normalidad

que exigen algunos métodos.

Para validar el método propuesto se desarrolla un análisis experimental sobre 534 series del alquiler de viviendas, para las 177 zonas de trabajo. Se han usado cinco técnicas de desagregación temporal clásicas: tres de tipo autorregresivo y dos no paramétricos. Los resultados muestran que las series obtenidas usando los métodos seleccionados por nuestra metodología ofrecen mejores indicadores de calidad que las series desagregadas, bien con un único método o una selección aleatoria. Además, por construcción, el método permite añadir nuevos criterios de evaluación de calidad o métodos de agregación, sin necesidad de cambiar el algoritmo. Por otra parte, no es necesario un calibrado manual del método, lo que lo hace aplicable a procesos de desagregación sobre un gran volumen de series.

En la primera parte del capítulo se hace una revisión de la literatura de las distintas técnicas de desagregación temporal de series, así como una mención a las recomendaciones oficiales de Eurostat (2015); la segunda parte describe la metodología aplicada a las series de precios generadas por el modelo hedónico final, y se seleccionan las series más adecuadas según el método propuesto; para, en la parte final, analizar los resultados obtenidos.

7.2 Desagregación temporal

Las series temporales son un instrumento clave para la medición y control de variables macroeconómicas (Eurostat, 2015), y aunque se disponga de información de forma exhaustiva para la cuestión de estudio, los datos pueden encontrarse con diferentes frecuencias temporales y niveles de agregación, lo que dificulta la integración de los datos. El estudio de esta problemática se ha venido realizado desde principios del siglo XX, y se han planteado un gran número de aproximaciones para resolverla (Moauro y Savio, 2005). Los métodos utilizados para homogeneizar las series de tiempo de baja y alta frecuencia, deben aplicar múltiples criterios como mantener el valor medio para todos los niveles de agregación o la ausencia de discontinuidades en la secuencia de valores.

Estas técnicas son ampliamente utilizadas por los institutos de estadística nacionales, como por ejemplo, Francia, Italia y otros países europeos, dónde se calculan las cifras trimestrales del Producto Interno Bruto (PIB) utilizando métodos de desagregación temporal (Sax y Steiner, 2013). En la Unión Europea, a través de Eurostat (2015) y el Sistema Estadístico Europeo (SEE) («European Statistical System», 2023), se han propuesto directrices para ayudar a los productores de datos a obtener series de alta frecuencia (trimestrales o mensuales) a partir de datos de baja frecuencia (anuales, bianuales, etc.).

Las aplicaciones prácticas se enmarcan en tres aspectos: desagregación temporal, evaluación comparativa y reconciliación. Las pautas dadas por Eurostat identifican las mejores prácticas para lograr tres objetivos: el primero, la armonización en todos los procesos nacionales; el segundo, la comparabilidad entre resultados; el tercero, la coherencia entre dominios y entre agregados con sus componentes.

El desglose se puede realizar sin ninguna serie de alta frecuencia, o incluso con más de una, denominadas indicadoras. En todo caso, cuando no hay una serie de indicadores, es posible desagregar temporalmente, si bien la precisión de la serie de alta frecuencia resultante será menor. Todos los métodos de desagregación aseguran que una función de agregación, sea la suma, el promedio, el primer o el último valor de la serie de alta frecuencia, guarda consistencia con la serie original. Estos métodos pueden hacer frente a situaciones en las que la alta frecuencia es un múltiplo entero de la baja frecuencia (por ejemplo, años a trimestres, semanas a días), pero no con frecuencias irregulares (por ejemplo, semanas a meses).

En las últimas décadas, debido a la mejora en las capacidades de proceso de información, se ha desarrollando un gran número de aproximaciones para hacer la desagregación (Dagum y Cholette, 2006a; Quilis, 2018), tales como los basados en: modelos de regresión, aproximaciones no paramétricas, basadas en ondas, mediante modelos bayesianos o redes neuronales, lo que plantea la dificultad de seleccionar el método adecuado.

7.2.1 Métodos de desagregación

Existen varias aproximaciones para la adaptación del dato de baja frecuencia a alta frecuencia en series temporales: el proceso de desagregación temporal propiamente dicho, la interpolación, el benchmarking o evaluación comparativa y la reconciliación de series. El objetivo final de cada técnica es combinar las series disponibles para transformar la serie de baja frecuencia a una de alta frecuencia, manteniendo la consistencia entre las mismas.

El benchmarking, o evaluación comparativa, pretende ajustar las discrepancias entre información estadística relacionada. Históricamente, se ha utilizado para la desagregación de magnitudes macroeconómicas anuales a frecuencias trimestrales, usando para ello otras medidas relacionadas con frecuencia trimestral[2].

[2]Los enfoques clásicos recogen un buen número de técnicas de desagregación temporal (Boot y Feibes, 1967; Cholette, 1984; Cholette y Dagum, 1994; Chow y Lin, 1971; Denton, 1971; Di Fonzo y Filosa, 1987; Fernandez, 1981; Guerrero y Martínez, 1995; Hillmer y Trabelsi, 1987; Lisman y Sandee, 1964; Litterman, 1983; Vangrevelinghe, 1966; Zani, 1970), para una revisión en detalle de

Las técnicas denominadas de reconciliación (Dagum y Cholette, 2006b) se utilizan para eliminar las discrepancias entre las series de alta y baja frecuencia, por ejemplo, las existentes entre una serie nacional de producción industrial procedente de un organismo público y las series individuales generada por una de las regiones de dicho país. Cuando las series desagregadas se combinan, es posible que su media no coincida con la serie original, en este caso, se requiere llevar a cabo un proceso de reconciliación. Las primeras propuestas de solución datan de la década de los 40 del siglo XX (Deming y Stephan, 1940), cuya evalución se produjo gracias al uso de los ordenadores para el procesamiento de información. Entre estos estudios encontramos, el de Cholette (1988), Chen y Dagum (1997), y Di Fonzo (2002).

Posteriormente, se publicaron varios artículos especialmente relevantes, tanto sobre los métodos de benchmarking como los de reconciliación. Di Fonzo y Marini (2005) resolvieron varios problemas de agregación de series y reconciliación mediante el principio de preservación de movimiento propuesto por Denton (1971). Posteriormente, los mismos autores desarrollaron un estudio empírico donde estudian las diferencias entre el método Denton Modificado con respecto a Causey Trager (Di Fonzo y Marini, 2011; Fonzo y Marini, 2013). Por otra parte, Daalmans y Di Fonzo (2014) analizaron las relaciones entre los métodos de mantenimiento de la tasa de crecimiento (Causey y Trager, 1981) y las primeras diferencias (Denton, 1971). Daalmans *et al.* (2018) compararon ambos métodos, y propusieron dos mejoras sobre el método Causey Trager que solucionan su irreversibilidad temporal y la singularidad de su función objetivo. Una de las contribuciones más recientes son las basadas en ondas (*wavelet*), que Davies *et al*. (2015), al compararlo con Denton y Cholette-Dagum, confirmaron su mejor comportamiento ante valores atípicos. Más recientemente, Sayal *et al.* (2017) introducen una evolución de este método.

La interpolación genera los valores a través del muestreo, en un intervalo temporal, de la serie original (Dagum y Cholette, 2006b; Eurostat, 2015). Los métodos más habituales son: la interpolación cúbica, por splines, o regresión local. La extrapolación, en cambio, calcula valores de una serie de tiempo para puntos de tiempo que no han sido muestreados y están fuera del intervalo de tiempo de la serie original. Entre estos métodos se encuentran los modelos ARIMA, suavizado exponencial, y los basados en regresión univariantes y multivariantes (Moauro y Savio, 2005).

De forma alternativa a las aproximaciones clásicas, se encuentran los métodos bayesianos, originalmente propuestos por Alba (1988), actualizados más tarde por

otros métodos clásicos véase (Miralles, 1997)

Rojo (2005, 2017). Estos se basan en encontrar la relación estadística entre el indicador y la serie a construir, y tiene como ventaja que favorece la regularidad de las series generadas, ofreciendo resultados maś estables aún teniendo series indicadoras de alta volatilidad. El uso de aprendizaje automático es más reciente y se ha aplicado a la desagregación de series no ecónomicas (Guyet *et al.*, 2022; Katranji *et al.*, 2016; Scher y Peßenteiner, 2021), a excepción del caso de Zaier y Abed (2014) que aplica redes neuronales a la desagregación de PIB de Estados Unidos.

El software disponible para esta cuestión ha estado restringido históricamente a las agencias estadísticas, donde los más habituales eran: ECOTRIM (Barcellan, 1994), BENCH de la oficina de Estadística de Canada, (Cholette y Dagum, 1994), o la librería MATLAB (Quilis, 2002). Esta situación ha cambiado en los últimos años, con la publicación por Sax y Steiner (2013) de un paquete de acceso libre en los lenguajes R y python.

El Anexo 7a de este capítulo, se describen en profundidad los cinco métodos de desagregación temporal aplicados en la metodología: Denton (1971), Denton y Cholette (Cholette, 1984), Chow-Lin (1971), Causey-Trager (1981) y Litterman (1983).

7.2.2 Consideraciones adicionales

Eurostat (2015) recomienda considerar tres criterios en el proceso de desagregación temporal de varias series de baja frecuencia, relacionadas con otras auxiliares de alta frecuencia: el primero, asegurar la calidad de los indicadores de baja frecuencia (en varios niveles de desagregación); el segundo, la disponibilidad y calidad de indicadores de alta frecuencia en varios niveles de desagregación; y el tercero, considerar los aspectos legales y necesidades de desagregación para los usuarios, y las consideraciones estadísticas de la calidad del resultado en los diferentes niveles de desagregación.

La estrategia de estimación está estrictamente relacionada con la identificación del mejor nivel de desagregación temporal. Cuando se deben estimar múltiples series vinculadas por restricciones de agregación, existen varias alternativas posibles: la primera, calcular por separado la variable agregada y sus componentes (método directo), aunque esto no garantiza el cumplimiento de las restricciones de agregación contemporánea; la segunda, el uso del método directo con técnicas de reconciliación, que aseguren el cumplimiento de las limitaciones contemporáneas; y el tercero, calcular los componentes individuales y derivar el valor final agregando los componentes estimados (método indirecto).

Para las tres alternativas anteriores, no existe un razonamiento teórico ni evidencia empírica a favor de un enfoque en particular. El directo es el preferido cuando hay co-movimientos entre agregados y componentes, y cuando la calidad de los componentes no es homogénea, lo cual es habitual a un nivel muy desagregado, el indirecto es la mejor opción si la calidad de estos últimos es lo suficientemente alta.

Por otra parte, no se debe obviar la importancia de los valores atípicos, por su impacto en la calidad del resultado de la desagregación temporal, evaluación comparativa y conciliación. Ante estos casos, es fundamental estudiar las causas de los valores atípicos para poder decidir el mejor tratamiento a aplicar. Si los valores atípicos aparecen en la serie indicadora solo como un único valor extremo (valores atípicos aditivos), y parecen originarse por un error en la fuente, es aconsejable corregirlos antes de ejecutar procedimientos de desagregación.

Si el valor atípico aparece tanto en el indicador de alta frecuencia como en la variable objetivo de baja frecuencia, y son explicables estadística o económicamente, deben modelarse durante el proceso de estimación o eliminarse antes de su inicio y reintroducirse al final del proceso. Finalmente, si aparece más de un valor atípico estructural en unos pocos indicadores, que reflejen cambios en el proceso de producción y que no se recogen en la variable original, se aplicaría un proceso de *winsorizacion*[3] de las altas frecuencias en las series indicadoras.

Además, es habitual encontrar índices de precios como series encadenadas, por ejemplo, en el IPV del INE (2016a), o en el Manual de Cuentas Nacionales de Eurostat (2013). En estos casos, Eutostat (2015) recomienda trabajar con series no encadenadas, aunque deja abierta la decisión en función de cada caso. El motivo es que una serie de precios encadenada no es, en sentido estricto, una serie temporal coherente, y por tanto, el encadenamiento puede distorsionar la naturaleza de los valores en los cambios de periodo (por ejemplo por el método de superposición trimestral).

[3] La *winsorización* es un proceso que limita una serie de valores superiormente a un valor máximo, e inferiormente a un valor mínimo.

7.3 Metodología

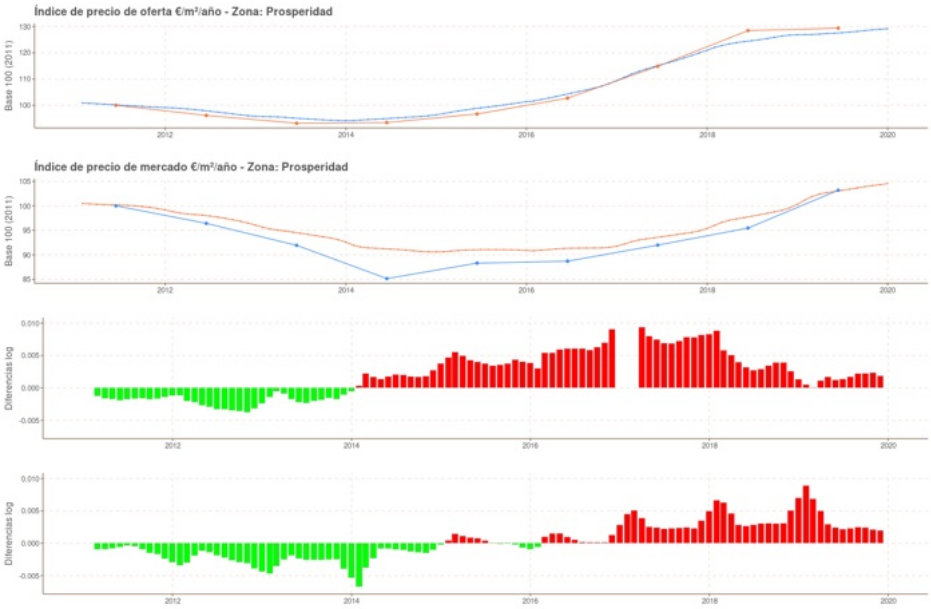
El método propuesto consiste en seleccionar automáticamente el modelo de desagregación temporal más adecuado para cada serie de precios. El proceso consiste en dos pasos: el primero selecciona, en cada zona geográfica, las series indicadoras a utilizar; el segundo desagrega con los 5 métodos métodos distintos las series de baja frecuencia de cada zona, y toma aquella serie con mayor verosimilitud de ser la óptima (en términos de calidad).

Esta aproximación resuelve los dos problemas fundamentales para la desagregación temporal, que remarca Quilis (2018), y que son: por una parte, disponer de mecanismos robustos de comparativa entre métodos; y por otra, encontrar herramientas que reduzcan la complejidad operativa de los métodos estadísticos habituales.

7.3.1 Paso 1: Selección de series indicadoras

Se parte de las series de precios generadas anuales y mensuales por el modelo final. En la Figura 7.2 se muestra un ejemplo para el barrio Prosperidad (Madrid)[4], en la cual se puede comprobar que las series desagregadas no cumplen la condición de que su media se corresponda al precio medio anual.

Figura 7.2. Series originales (oferta y alquiler) con diferencias logarítmicas mensuales. Barrio Prosperidad (Madrid)

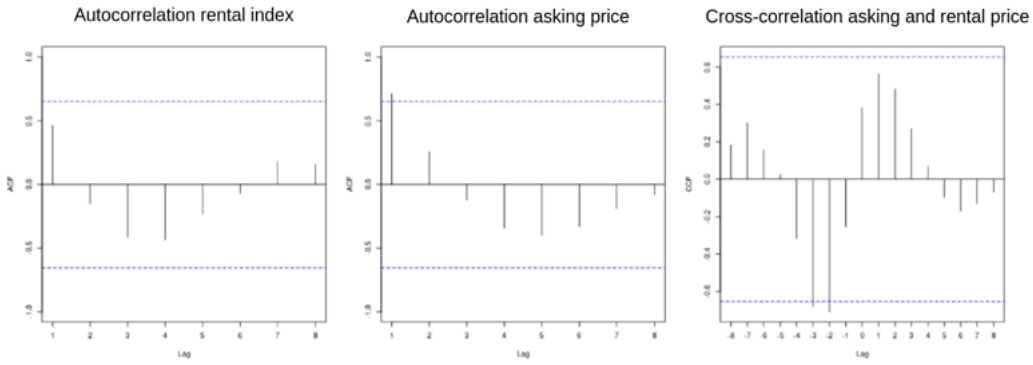


Fuente: elaboración propia.

[4] Se toma el barrio de Prosperidad por ser una zona mercado de alta demanda y numerosa muestra, en la que se aprecian diferencias entre los datos mensuales y anuales generados por los modelos de correspondencia. En particular, las medias de los datos mensuales no se corresponden con los datos anuales. La información de todas las series está disponible en los enlaces del Anexo I.

No obstante, existe una alta la coherencia temporal entre las series de alta y las de baja frecuencia, por tanto se han tomado las series mensuales (series indicadoras) como base para la desagregación. Tras un análisis preliminar, se observa la existencia de una correlación alta entre la serie de oferta y alquiler, aunque con un desfase temporal, como muestra la Figura 7.3. Las autocorrelaciones de las series anuales de oferta y alquiler, y la correlación cruzada entre alquiler y oferta, demuestran que la serie de alquiler va retrasada con respecto a la de oferta. La interpretación económica del fenómeno anterior es que la oferta actúa de forma adelantada al mercado, con diferencias en función de las características del mercado (Kokot y Bas, 2015), principalmente en función del grado de absorción de la oferta (Galesi *et al.*, 2020).

Figura 7.3. Autocorrelación series anuales de alquiler, de series en oferta y cruzada oferta y alquiler. Pacífico (Madrid)



Fuente: elaboración propia.

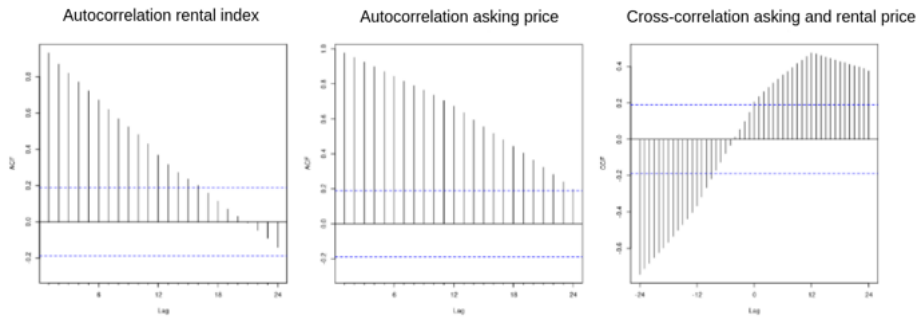
El grado de absorción, o liquidez del mercado, se puede estimar bien a través de la tasa de reposición del *stock* en oferta, o por su reflejo en la tensión de la demanda, medida como número de contactos medios mensuales de los anuncios de una zona[5]. La serie de alquiler muestra un retraso de entre uno y dos años respecto de la serie de alquiler[6] .

La Figura 7.4 muestra que las autocorrelaciones entre las series mensuales son muy fuertes en los meses contiguos, como consecuencia de la mayor influencia del componente de tendencia. La autocorrelación cruzada refleja ciclos más largos en la oferta que en el alquiler, con una serie de oferta que va, aproximadamente, un año por delante de la de alquiler.

[5]Este dato procede del portal inmobiliario Idealista. Un contacto es un dato recogido por la plataforma que se produce cuando un demandante de vivienda envía un mensaje al propietario indicando que tiene interés en alquilarla o adquirirla.

[6]Este fenómeno depende de la zona y de la capacidad de abosorción y reposición del *stock* de viviendas de alquiler por nuevos contratos, se puede decir que el dinamismo del mercado guarda una relación con el retraso entre series.

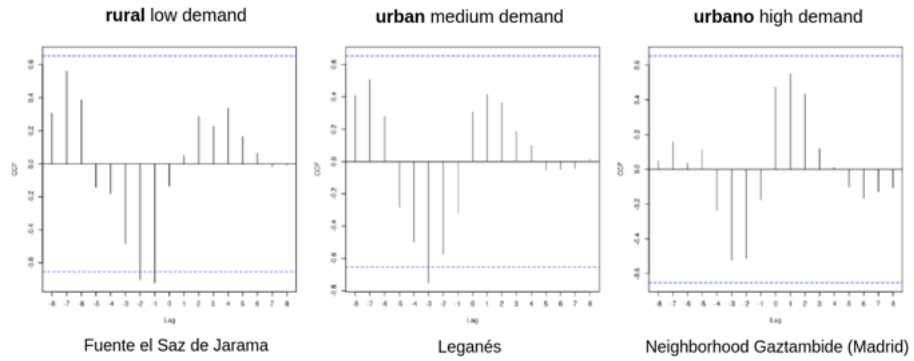
Figura 7.4. Autocorrelación series mensuales de alquiler, de series en oferta y cruzada oferta y alquiler. Pacífico (Madrid)



Fuente: elaboración propia.

A tenor de lo comentado anteriormente, las dinámicas de mercado parecen indicar que aquella áreas menos activas reaccionan más lentamente ante cambios en la oferta. Para corroborar la hipótesis anterior se han tomado tres mercados representativos de áreas clasificadas según su dinamismo inmobiliario: Área de Fuente el Saz de Jarama (zona rural, de baja demanda y, por tanto, poco activa); el municipio de Leganés (zona urbana de demanda media, con un dinamismo mediano); y el barrio de Gaztambide de Madrid (zona urbana de alta demanda que representa un mercado con alta liquidez). Dichas zonas, representadas en la Figura 7.4, muestran que las correlaciones cruzadas en la zona rural tienen un rango mucho más amplio que las zonas urbanas. Por otra parte, Gaztambide tiene un rango más estrecho, lo que significa que el mercado de alquiler reacciona más rápidamente ante el comportamiento de la oferta. Se puede, por tanto, confirmar la hipótesis de que la longitud del retraso es inversamente proporcional al nivel de actividad del submercado inmobiliario. Por tanto, la estructura temporal de la serie dependerá de la dinámica de oferta y demanda de la zona y determinará la longitud del adelanto oferta-mercado y la excentricidad de los cambios.

Figura 7.5. Correlación cruzada de series mensuales de oferta y alquiler: Zonas Fuente del Saz, Barrio Gaztambide y Leganés



Fuente: elaboración propia.

Las 177 localizaciones cuentan con dos series temporales de baja frecuencia (anual), una con el índice de precios de compra y otra con el precio del alquiler. Existe una gran diversidad de zonas ya que se incluyen tanto con barrios en el centro de la ciudad de Madrid como municipios de tipo rural de la comunidad de Madrid. Como *a priori* no se conoce cual es el mejor método y cada serie tiene una naturaleza diferente, se opta por un proceso de selección automática del método a aplicar.

La estrategia a seguir, para cada serie, asume de que el método a seleccionar es el que ofrezca la mayor de verosimilitud para desarrollar una desagregación temporal óptima. Dicha verosimilitud atiende al comportamiento estructural de la serie, la validez de los datos, la consistencia entre las series de alta y baja frecuencia, y la coherencia entre la serie generada y su serie indicadora.

Se crean un total del 10 series indicadoras para las series zonales, formadas por 5 grupos de zonas de Madrid y otros 5 para el resto de la provincia. Dichos grupos se construyen en función del nivel de demanda[7] con el algoritmo K-Medias (Lloyd, 1982). El tamaño de cada grupo se recoge en la Tabla 7.1, observándose que en la capital hay una mayor concentración de zonas de los grupos 2 y 3 (demanda media), mientras que en el resto de la Comunidad, las zonas están distribuidas de una forma más equitativa entre grupos.

Tabla 7.1. Número de zonas contenidas en cada grupo

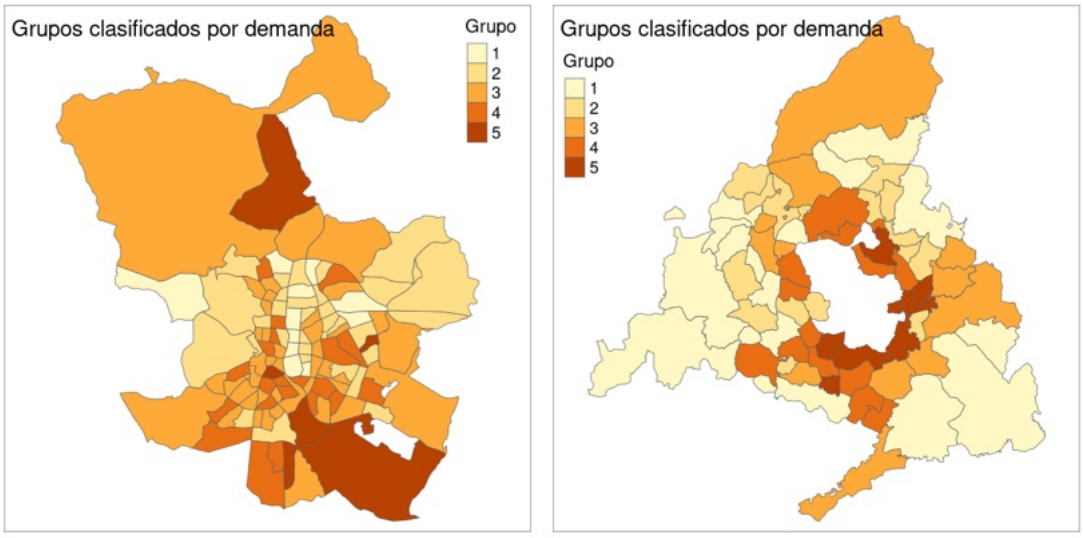
Área	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Madrid	11	32	42	21	6
Resto CAM	17	15	12	13	8

Fuente: elaboración propia

El desglose zonal de grupos, mostrado en la Figura 7.6, refleja que en la capital los grupos se encuentran bastante repartidos geográficamente, mientras que el resto de la Comunidad aquellos con mayor demanda se agrupan de forma concéntrica alrededor de la capital.

[7] Medida como el número de contactos medios por anuncio al mes.

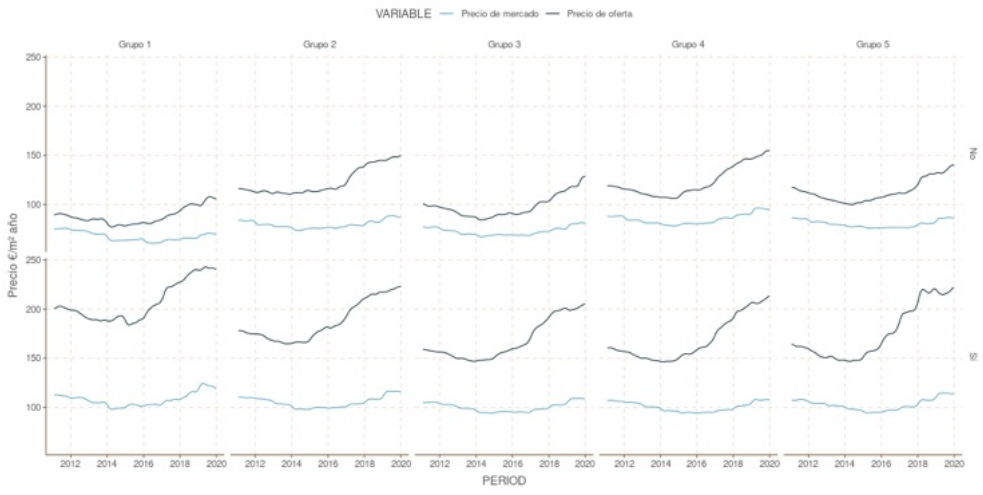
Figura 7.6. Grupos de zonas para la creación de series indicadoras de mercado



Fuente: elaboración propia.

Sobre las 10 series mensuales indicadoras se aplica una media móvil triangular[8] con $N = 3$, cuyo resultado se representa en la Figura 7.7. Se aprecian que, a pesar del suavizado, las transiciones anuales son relativamente bruscas.

Figura 7.7. Series indicadoras sin procesar utilizadas para el desglose zonal, grupos Madrid y resto de provincia



Fuente: elaboración propia.

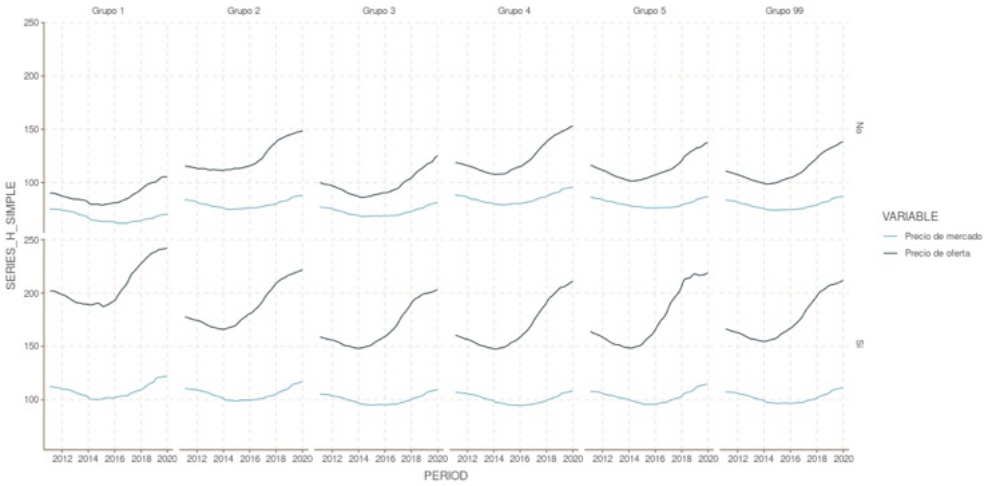
Las variaciones abruptas entre los meses finales e iniciales de año son consecuencia de la falta de capacidad de los modelos hedónicos de incorporar adecuadamente las variaciones intermensuales entre enero y diciembre. La causa se debe a que los árboles que especifican el tiempo a través de variables dummy

[8] Es una media móvil cuyo dato en momento t se calcula como la media de los datos originales en los periodos: $t-1$, t y $t+1$.

de periodo, tienden a infraestimar la influencia de los componentes de tendencia y estacionalidad. Como consecuencia, ofrecen valores cercanos a la media la media anual de los precios, ignorando la componente temporal.

Para mitigar el efecto anterior, se descomponen las series indicadoras originales promediándolas con series de interpoladas[9]. El resultado suaviza los cambios entre años como se aprecia en la Figura 7.8.

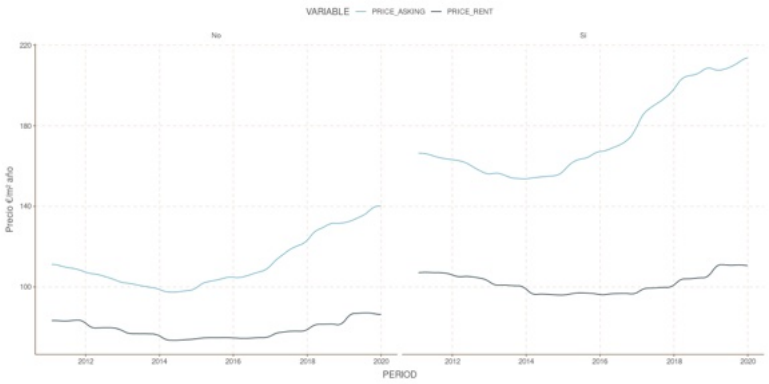
Figura 7.8. Series indicadoras finales, Madrid y resto de Comunidad



Fuente: elaboración propia.

Para los índices de tipo funcional se utiliza un desglose por capital o resto de provincia análogo, como muestra la Figura 7.9.

Figura 7.9. Series indicadoras utilizadas para el desglose funcional, grupos Madrid y resto de provincia



Fuente: elaboración propia.

[9]Las series interpoladas se construyen mediante la interpolación mensual de los valores anuales.

7.3.2 Paso 2: Selección de las mejores series

En esta etapa del proceso, se selecciona la mejor serie a partir de un conjunto de pruebas utilizando diferentes métodos. Todos ellos emplean un enfoque univariante de desagregación temporal[10] con los 5 métodos mostrados en la Tabla 7.2. Los cuales se encuentran entre las técnicas más habituales en las agencias estadísticas[11] (Quilis, 2018) (Moauro y Savio, 2005). El método Chow-Lin, utiliza la variante propuesta por Silva y Cardoso (2001), dónde el parámetro autorregresivo ρ se estima automáticamente . Se descarta el método Denton original por su inestabilidad en los primeros valores de las series (Cholette y Dagum, 1994).

Tabla 7.2. Métodos de desagregación temporal de series temporales

Método	**Tipo**	**Notas**
Chow-Lin Max Log	Regresión	Basado en Chow-Lin
Litterman Max Log	Regresión	Basado en Litterman
Dynamic	Regresión	Basado en Litterman
Denton-Cholette	Denton	Implementación Denton-Cholette
Causey-Trager	No paramétrico	Basado en el método Causey-Trager

Fuente: elaboración propia

La efectividad de cada metodología puede fluctuar dependiendo de la especificidad inherente a las series de datos en consideración. Mientras que es viable optimizar la parametrización de cada modelo para alcanzar un mejor rendimiento en cada caso, los manuales suministrados por los institutos estadísticos no ofrecen un criterio cuantitativo unificado para la elección del método que se adecúe de modo más pertinente. Por lo general, la decisión se basa en la opinión de un experto (Moauro y Savio, 2005; Quilis, 2018). Sin embargo, un proceso manual como este no permite la desagregación eficiente ante un gran número de series con características diversas, ya que requiere la intervención de muchas personas y, la aproximación de desagregación empleada podría no ser la más apropiada, si se utiliza un solo modelo para todos los casos.

Se propone un método cuantitativo objetivo que imita el criterio del experto, el cual consiste en seleccionar la serie que tenga un comportamiento más similar al de la serie de referencia y que, además, presente unas métricas estructurales cercanas al valor óptimo.

[10] Se ha utilizado el paquete R *tempdisagg* (Sax y Steiner, 2013).

[11] La popularidad de estos métodos se debe a su conexión los modelos lineales generales de econometría. Además, de su relativa simplicidad que facilita la comunicación con los usuarios finales.

Se define un estimador de máxima verosimilitud ($\mathcal{L}$), que selecciona la serie cuya medida de verosimilitud θ es máxima para los parámetros de la misma. El valor final se calcula como la probabilidad de un conjunto de criterios de calidad c, calculados sobre la serie desagregada H con un método m dado. Para facilitar su cálculo, se asume la independencia los sucesos de los criterios, de manera que la verosimilitud se estima como el producto de las probabilidades individuales, según:

$$\mathcal{L}(\theta|\hat{H^m}) = \prod_{c=1}^{n} p(\hat{H_c^m} \mid \theta_c) \quad [7.1]$$

donde $\hat{H^m}$ es la serie mensual estimada con el método m, y θ_c una medida de verosimilitud según el criterio de calidad c.

Para facilitar la comprensión del método, la expresión anterior se puede calcula, para cada criterio de calidad, la probabilidad de que dicha que serie desagregada H pertenezca a la distribución óptima de valores óptima (según dicho criterio).

Finalmente, se seleccionará el método de desagregación m, de entre los distintos métodos candidatos, que hace máxima la expresión de verosimilitud de la expresión 7.1, y cuya definición recoge la siguiente expresión:

$$\hat{\theta}(\hat{H})_{mle} = arg\ max_{\theta \in \Theta}\ \mathcal{L}(\theta \mid \hat{H^m}),\ \forall m \in m_1, ..., m_k \quad [7.2]$$

Que transformada a verosimilitudes logarítmicas sería:

$$log\ \hat{\theta}(\hat{H})_{mle} = arg\ max_{\theta \in \Theta}\ \sum_{i=1}^{n} log\ p(\hat{H_c^m} \mid \theta_c),\ \forall m \in m_1, ..., m_k \quad [7.3]$$

En el cálculo de las probabilidades de cada criterio se asume que siguen una distribución normal-logarítmica, por lo que se estiman los parámetros de una distribución de dicho tipo[12] que se más se ajuste a los datos originales. Para aumentar la robustez de la estimación, se eliminan los valores extremos de cada variable (menores del percentil 1 y mayores del percentil 99).

7.3.3 Criterios de evaluación de la calidad de las series

La verosimilitud se estima para cinco criterios de calidad que forman parte de los propuestos por Chen y Andrews (2008) y que se miden a través de cinco métricas correspondientes. Estos criterios reflejan el buen comportamiento de una serie temporal:

[12] A través del paquete R *fitdistrplus* (Delignette-Muller *et al.*, 2015).

- Deben cumplirse los requisitos de agregación: en nuestro caso, se impone que la media de las series desagregadas coincida con el valor de la serie agregada. El nivel de discrepancia se calcula como error cuadrático[13].
- No debe haber variaciones bruscas entre un año y otro. Como afirma Hood et al. (2005) en el estudio comparativo de procesos de benchmarking, la homogeneidad a lo largo del año no es consistente, lo que puede causar distorsiones importantes en ciertos métodos. Para evaluar este cambio, Chen y Andrews (2008) prefieren medir la media de las variaciones entre noviembre y febrero, y compararla con la media de las variaciones entre marzo y octubre.
- Los movimientos de corto plazo del indicador deben mantenerse lo máximo posible: para ello se compara la discrepancia entre las diferencias logarítmicas de la serie indicadora con la serie generada.
- La serie estimada no debe mostrar cambios bruscos al inicio o fin de la serie, situación habitual en métodos como el Denton original.
- La serie no debe ofrecer valores fuera de rango, en particular, los valores nunca deben ser negativos.

En consecuencia, se definen cinco métricas para evaluar el cumplimiento de los criterios anteriores, recogidas en la Tabla 7.3. La variación intraperiodo se calcula como diferencias logarítmicas en valor absoluto, y permiten comparar los valores de cambio entre diferentes zonas. Siendo $\hat{H}_t^{m,z}$ el precio medio estimado [14] por el modelo m de una zona z para un periodo mensual t, y la diferencia logarítmica:

$$\Delta_{log}(\hat{H}_{t,t-1}^{m,z}) = |log(\hat{H}_{t-1}^{m,z}) - log(\hat{H}_t^{m,z})| \quad [7.4]$$

Tabla 7.3. Métricas de evaluación de series

Requisito	**Métrica**
Agregación MSPE	Media mensual con valor anual (MSE)
Cambios inter-anuales	Variación máxima frontera anual (VF)
Movimientos a corto plazo	Media de las variaciones mensuales (VM)
Cambios bruscos inicio y fin	Ratio variación periodos extremos / mediana (REP)
Positividad de la serie	0 si hay negativos, 1 en caso contrario (POS)

Fuente: elaboración propia

A continuación se describe el método de cálculo de las 5 métricas, siendo h_t la serie de baja frecuencia original, $\hat{H}_t$ la serie estimada de alta frecuencia e I_t la

[13] El error cuadrático se calcula como la diferencia al cuadrado entre el dato anual y la media de las series desagregadas.

[14] H representa el precio real del alquiler y $\hat{H}$ el estimado por el modelo.

serie indicadora de alta frecuencia utilizada.

- *MSE:* error cuadrático medio tipificado de la media anual de alquiler y la serie anual original, y expresada en porcentaje para hacerla comparable entre zonas, definida a continuación:

$$MSE(\hat{H}^{m,z}) = \left[\mu(H^{m,z}) - \mu(\hat{H}^{m,z})\right]^2 \quad [7.5]$$

  donde $\mu(H_t^z)$ es la media anual de la serie para la zona z y el momento t, y la $\mu(H_t^{m,z})$ la media mensual.
- *VF*: máxima variación de frontera anual, medida como la máxima diferencia logarítmica en valor absoluto entre el mes de diciembre y el mes de enero del año siguiente, de la serie temporal de valores mensuales, según:

$$VF(\hat{H}^{m,z}) = max\ |log\left(\frac{\hat{H}_{y,12}^{m,z}}{\hat{H}_{y+1,1}^{m,z}}\right)|, \forall y \in [y_1..y_n] \quad [7.6]$$

  donde $\hat{H}_{y,12}^{m,z}$ es el valor de la serie mensual para diciembre y $\hat{H}_{y,1}^{m,z}$ el dato para enero.
- *VM*: variación media calculada como la media de las diferencias logarítmicas de la serie, calculada como:

$$VM(\hat{H}^{m,z}) = \frac{1}{N-1}\sum_{t=1}^{N-1} |log(\hat{H}_{t+1}^{m,z}) - log(\hat{H}_t^{m,z})| \quad [7.7]$$

  donde $\hat{H}_t^{m,z}$ es el valor de la serie candidata en la zona z y el mes t, y $\hat{H}_{t+1}^{m,z}$ el correspondiente valor para el mes siguiente.
- *REP*: mide la diferencia de la variaciones de los periodos inicial $t = 1$ y final $t = n$.

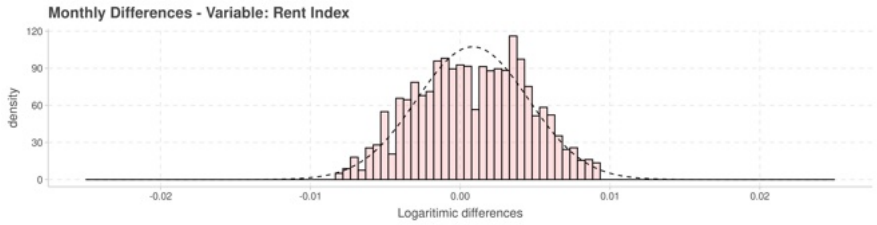
$$REP(\hat{H}^z) = \frac{argmax\left[\Delta_{log}(\hat{H}_{2,1}^z), \Delta_{log}(\hat{H}_{n,n-1}^z)\right]}{med\{\Delta_{log}(\hat{H}_{t,t-1}^z)\}} \quad [7.8]$$

- *POS*: de positividad, calculada como un indicador con valor 1 si la serie tiene todos los valores positivos, y 0 si tiene alun valor negativo. Esta medida se incluye porque algunos métodos de regresión pueden ofrecer valores negativos.

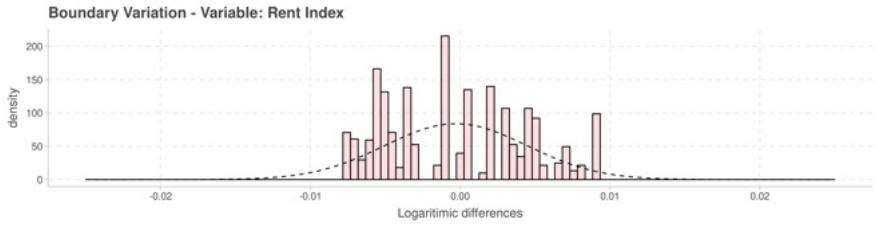
Para el caso de la distribución de variaciones logarítmicas mes a mes del alquiler, se obtiene una función de distribución normal estimada $\mathcal{N}(0, 0.065)$. La Figura 7.10 muestra dos de las funciones estimadas. En el caso de las diferencias mensuales, la aproximación de la distribución normal tiene un buen ajuste, en cambio, para la

variación frontera la forma es menos asimilable a una distribución normal. En ese caso, se acepta al ser más restrictiva antes variaciones mayores en valor absoluto y favorecerá transiciones año a año menores.

Figura 7.10. Funciones de densidad de la función de verosimilitud



(a) Variación de precios mensuales



(b) Variación de precios frontera

Fuente: elaboración propia.

Para el caso particular de la métrica *POS*, la función de densidad utilizada para estimar la verosimilitud no es una normal, sino que toma dos valores: para 1 (todos positivos), la verosimilitud es 1; y para 0, el valor es 10^{-10}.

7.4 Resultados

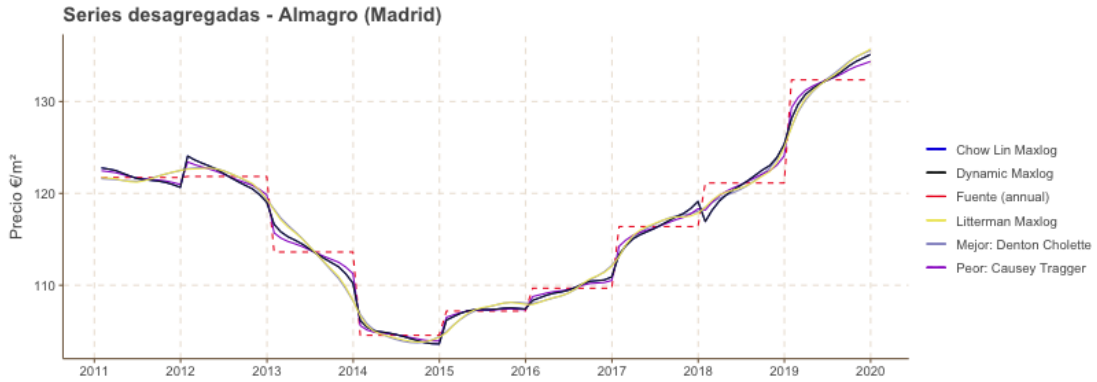
El método propuesto se ha evaluado sobre las series de alquiler de las 177 áreas de estudio, en base a tres aspectos:

- Comparativa de las métricas de calidad de los cinco modelos.
- Mejora del método comparado con una selección aleatoria de modelo.
- Diferencias entre modelos autorregresivos y no autorregresivos.
- Influencia del criterio de calidad.

Los tres ángulos anteriores se examinan tanto para las series candidatas como para la serie seleccionada de cada zona. Para su mejor comprensión, a continuación, se presenta un ejemplo representativo de todas ellas en la Figura 7.11, correspondiente al caso de un barrio madrileño con un comportamiento medio típico, como es el caso de Almagro (para ver todos los ejemplos, consúltese el Anexo I). La serie discontinua de color naranja ilustra los valores de la serie anual, que se busca desagregar, mientras que la serie discontinua de mayor

grosor y en color negro representa la serie óptima seleccionada mediante el método propuesto en la sección 7.3.2. El resto de las series, mostradas en diferentes colores, corresponden a los resultados de los 4 métodos que no fueron considerados óptimos.

Figura 7.11. Desagregación mensual del precio del alquiler. Zona: Barrio de Almagro (Madrid)



Fuente: elaboración propia.

De los cinco criterios de evaluación presentados en la metodología se ha decidido excluir en esta sección los resultados del de ajuste por MSE, ya que todos los métodos analizados presentan diferencias extremadamente bajas en todos los casos estudiados, siendo estas inferiores a 10^{-10}. Por consiguiente, la evaluación del ajuste basado en el MSE no proporciona información valiosa ni relevante para el análisis en curso. Por este motivo se ha optado por centrar nuestra atención en otros criterios que permitan diferenciar y comparar más eficazmente los distintos métodos aplicados en la investigación.

7.4.1 Comparativa de métricas de calidad

La Tabla 7.4, muestra la distribución de los métodos seleccionados con el método aplicado para las series de alquiler desagregadas. Para cada uno, se representa la probabilidad media por criterio y la verosimilitud media total, se observa que el método más seleccionado es el método autorregresivo Litterman Maxlog, seguido por Dynamic Maxlog, que además ofrece mayor probabilidad para los criterios de variación frontera y mensuales.

Los métodos autorregresivos son los que acaparan más del 95% de las series finales, en detrimento de Denton y Causey-Trager, que muestran peor comportamiento en los cambios de año. La probabilidad del 100% en el criterio de positividad, indica que ninguna serie ganadora[15] ofrece valores negativos, y

[15] El término de "serie ganadora" se refiere a aquella que se selecciona por el modelo, por considerarla

por tanto no será un criterio discriminante para comparar métodos.

Tabla 7.4. Probabilidad y verosimilitud para las series seleccionadas, desglosadas por método ganador.

Método	% Casos	Prob. VF	Prob. VM	Prob. POS	Prob. final
Denton Cholette	34.7%	12.08%	17.75%	100.0%	2.27%
Dynamic Maxlog	30.2%	15.70%	17.13%	100.0%	2.96%
Litterman Maxlog	29.0%	17.45%	20.95%	100.0%	4.16%
Causey Tragger	3.3%	28.27%	17.26%	100.0%	5.83%
Chow Lin Maxlog	2.9%	18.11%	14.98%	100.0%	2.92%

Fuente: elaboración propia

Para el criterio de variaciones mensuales, en general, tal como se evidencia en la Tabla 7.5, los métodos con probabilidades medias más bajas están relacionados con valores de medidas de VF más extremos. Esto significa que, a mayor variación entre diciembre y enero, menor es la probabilidad de que dicho evento ocurra en la distribución de referencia.

Para el caso del método Causey Trager, se observa una probabilidad media alta pero con unos valores de variación frontera media intermedia similares al resto de casos. Este comportamiento se debe una la alta dispersión de los valores de VF para este método.

Tabla 7.5. Relación entre valores medios de variación frontera y probabilidad

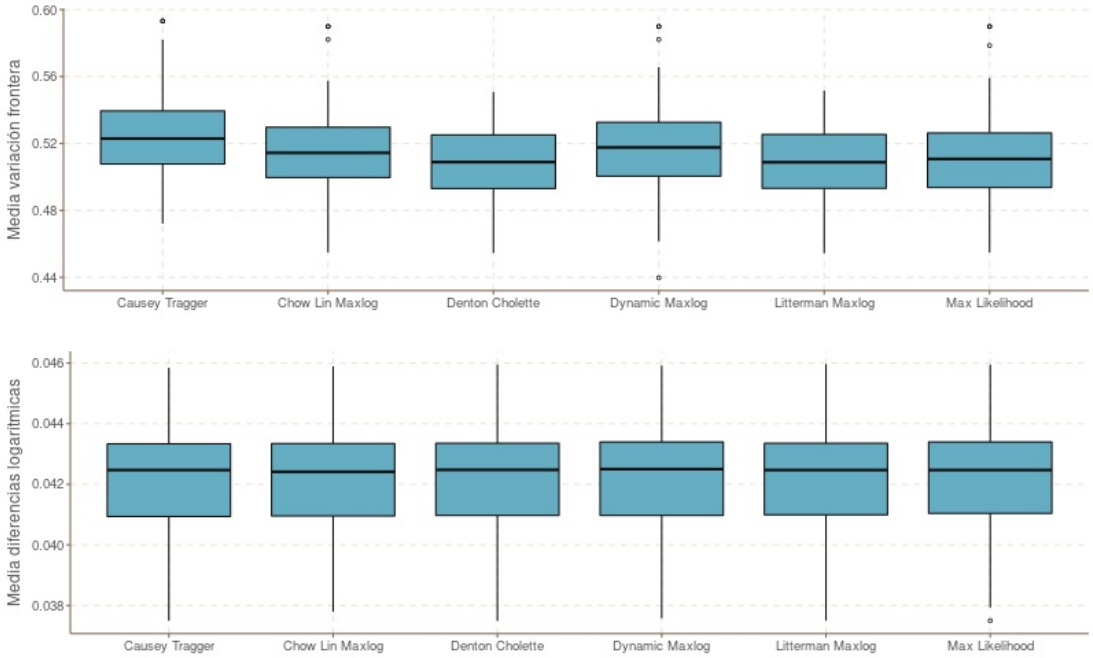
Método	% Casos	media VF	mediana VF	Stdev. VF	Prob. VF
Denton Cholette	34.7%	0.553	0.557	0.032	12.08%
Dynamic Maxlog	30.2%	0.548	0.549	0.023	15.70%
Litterman Maxlog	29.0%	0.550	0.558	0.035	17.45%
Causey Tragger	3.3%	0.563	0.557	0.026	28.27%
Chow Lin Maxlog	2.9%	0.560	0.569	0.038	18.11%

Fuente: elaboración propia

El estudio de los rangos intercuartílicos de los distintos métodos, como muestra la Figura 7.12, indica un mayor grado de variabilidad en el método Denton-Cholette, con la presencia de valores muy extremos en ambas medidas. El resto de los métodos ofrecen comportamientos similares, sin apreciarse diferencias notables en las series seleccionadas por el MLE.

óptima dentro de las 5 candidatas.

Figura 7.12. Comparación de métricas clave según los diferentes métodos



Fuente: elaboración propia.

En el análisis comparativo de diversos métodos aplicados a las series de alquiler desagregadas, el enfoque autorregresivo se posiciona como el preferido por el modelo, representando más del 95% de las series finales. Se observó, además, una propensión menor a seleccionar los métodos que ofrecen variaciones intermensuales extremas o diferencias abruptas entre enero y diciembre, como es el caso de Denton-Cholette y Causey-Trager, respectivamente.

7.4.2 Mejora incremental del método

Para evaluar cuanto mejor es una selección por máxima verosimilitud que un único método, se comparan las probabilidades y verosimilitud de las series generada por MLE contra una selección de un sólo método. El cual puede ser fijo o seleccionando de forma arbitraria para cada zona (aleatorio[16]).

Para el primer escenario, se comparan la probabilidad y para la medida VF de cómo funcionaría un proceso que seleccionara un método de desagregación aleatoria. Como se observa en la Tabla 7.6 de probabilidad, y la Tabla 7.7 de valores de VF, la selección aleatoria es notablemente peor. Con probabilidades de verosimilitud menores y peores indicadores de calidad, que si hubiéramos tomado el método recomendado por el criterio de máxima verosimilitud.

[16] Se denomina aleatorio, porque en este caso el método elige un método al azar para cada una de las zonas.

Tabla 7.6. Comparativa de probabilidades en series selecciondas por maxima verosimilitud contra una selección aleatoria

Método	Prob. VF	Prob. VM	Prob. POS	Prob. final
Máxima verosimilitud	37.65%	20.91%	100.00%	8.88%
Aleatorio	24.94%	16.79%	100.00%	5.64%

Fuente: elaboración propia

Tabla 7.7. Comparativa de valores en series selecciondas por maxima verosimilitud contra una selección aleatoria

Método	media VF	mediana VF	Stdev. VF
Máxima verosimilitud	0.511	0.511	0.023
Aleatorio	0.514	0.515	0.023

Fuente: elaboración propia

En el segundo caso, se estudia la mejora comparativa contra un único métodos. Se corrobora de nuevo la superioridad del método, como se aprecia en la Tabla 7.8. tanto la probabilidad como las medidas de VF y VM son claramente mejores en método de máxima verosimilitud. Se observa, además, que el uso de un único método inapropiado puede afectar seriamente a los resultados finales (ver caso de Causey-Trager).

Tabla 7.8. Comparativa de probabilidades en series selecciondas por maxima verosimilitud contra todos los métodos y una selección aleatoria

Método	Prob. VF	Prob. VM	Prob. POS	Prob. final
Máxima verosimilitud	37.65%	20.91%	100.00%	8.88%
Denton Cholette	32.38%	20.41%	100.00%	7.57%
Litterman Maxlog	30.01%	19.92%	100.00%	6.94%
Dynamic Maxlog	27.54%	15.54%	100.00%	6.45%
Aleatorio	24.94%	16.79%	100.00%	5.64%
Chow Lin Maxlog	23.40%	17.45%	100.00%	5.33%
Causey Tragger	11.36%	10.61%	100.00%	1.92%

Fuente: elaboración propia

7.4.3 Modelos autorregresivos contra no autorregresivos

En el análisis los métodos no autorregresivos, se confirma que los métodos Causey-Trager y Denton-Cholette muestran dificultades para suavizar las transiciones en el cambio de año, tal y como muestra la Figura 7.13. Esta situación se corrobora

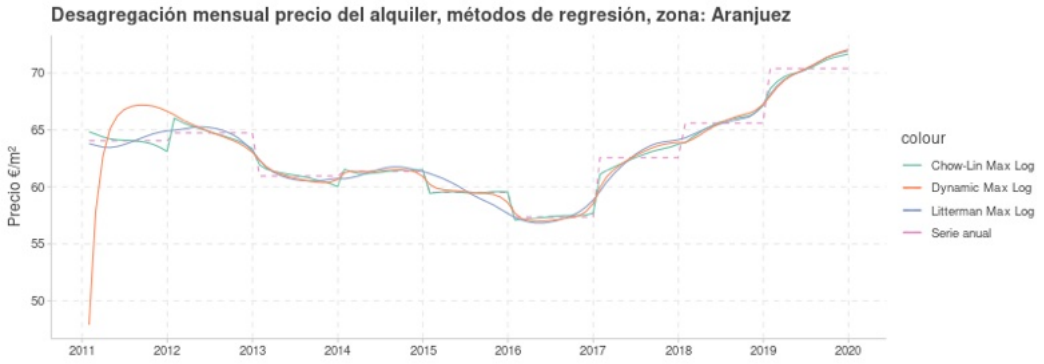
con una menor probabilidad de la variación frontera de los métodos como se observa en la Tabla 7.8.

Figura 7.13. Desagregación no paramétrica: Denton y Causey-Trager

Fuente: elaboración propia.

Los métodos de regresión ofrecen una mayor capacidad para suavizar las transiciones anuales, como muestra la Figura 7.14. En el método Dynamic Maxlog no es infrecuente encontrar valores extremos al inicio de las series como vemos en la propia figura. Si bien es cierto que no tanto como en el método Denton original, que se descartó por mostrar frecuentemente este patrón. Por contra, los de tipo autorregresivo muestran transiciones más suaves tanto en las fronteras, entre los meses de diciembre y enero, como en los cambios mes a mes.

Figura 7.14. Desagregación basados en regresión



Fuente: elaboración propia.

La ventaja de los métodos paramétricos es la posibilidad de estudiar la bondad del ajuste. Para el caso de estudio, tal y como muestra la Tabla 7.9, el ajuste de tanto Dynamic Maxlog y Chow-Lin en R^2 es muy alto, y, en cambio, para Litterman existe un valor sensiblemente más bajo. Las diferencias pueden estar motivadas por la mayor sensibilidad de este último método ante valores atípicos y series ruidosas,

además de que los dos primeros métodos se adaptan mejor a series sin un patrón de tendencia estable (véase Anexo 7a).

Tabla 7.9. Parámetros principales de las diferencias logarítmicas

Método	R^2
Dynamic Maxlog	87.68%
Chow Lin Maxlog	80.25%
Litterman Maxlog	65.69%

Fuente: elaboración propia

7.4.4 Influencia del criterio de calidad en la selección

Puesto que la decisión de seleccionar una serie está condicionada a la verosimilitud de la combinación de una serie de criterios, un cambio en el modo de agregarlos puede dar lugar a una selección totalmente diferente. Para evaluar el grado de sensibilidad de la misma para un caso concreto, se comparan los resultados en caso de seleccionar a través de un solo criterio.

En primer ejemplo, que usa solamente el error cuadrático de los valores y representado en la Figura 7.16, muestra como la mejor selección tiene un nivel bajo de variabilidad mensual, en el primer año replica la forma del escalón y las transiciones anuales de los primeros años son ligeramente abruptas.

Figura 7.15. Mejor y peor serie por MSE

Fuente: elaboración propia.

Para la variación frontera (Figura 7.16), se observa un comportamiento más suave a lo largo del tiempo, favorecienciendose la transición suave entre años.

Figura 7.16. Mejor y peor serie por verosimilitud por VF

Fuente: elaboración propia.

Finalmente, si el criterio son las variaciones intermensuales (Figura 7.17), se aprecia que la selección coincide con el caso anterior, lo que estaría asociado a que el criterio VF contiene parcialmente los requisitos del criterio VM.

Figura 7.17. Mejor y peor serie por VM

Fuente: elaboración propia.

En resumen, se constata la superioridad de los métodos autorregresivos sobre los que no lo son, para el caso de estudio. El motivo es la dificultad de los métodos Causey-Trager y Denton-Cholette para suavizar las transiciones en los cambios de año y los casos de alta variabilidad, ambas son cuestiones específicas de las series generadas por nuestro modelo y que se ligan a los problemas de estimación mencionados en los capítulos precedentes.

Por otra parte, los métodos autorregresivos también muestran diferencias debidas a las peculiaridades de las series de entrada. Siendo los métodos más sensibles al ruido en las series originales aquellos que pueden provocar más dificultad a ciertos métodos.

Finalmente, aunque las series generadas cualquiera de los métodos solucionan de una forma razonable el problema, el uso de una verosimilitud combinada en

base a distintos criterios ofrece un método más robusto ante las debilidades de un criterio en particular.

A lo largo de este capítulo se ha presentado una novedosa metodología para asistir el proceso de desagregación temporal. Las series resultantes aseguran la coherencia y la calidad temporal entre los distintos niveles de agregación, al igual que se ha conseguido la coherencia zonal en el método presentado en el capítulo anterior. En el siguiente capítulo, se analizan las series temporales generadas para los índices de precio de oferta y de mercado, atendiendo a sus características estructurales como de mercado.

Anexo 7a. Métodos de desagregación temporal

Método Denton

El método de benchmarking conocido como método Denton (Denton, 1971) o de primeras diferencias, tiene como objetivo minimizar los cambios en las series de alta frecuencia, en adelante indicadoras, al tiempo que cumple un conjunto de restricciones de evaluación comparativa.

Este método minimiza los cuadrados de las desviaciones, absolutas o relativas, de la serie original y de la serie indicadora, denominándose el parámetro h como el grado de diferencias. La técnica se basa en el principio de preservación del movimiento, sobre la que existen las modalidades aditiva y proporcional. En el caso de la versión aditiva, cuando $h = 0$ se minimiza la suma de los cuadrados de las diferencias entre la serie original y la indicadora, para $h = 1$ se minimiza la suma de las primeras diferencias, para $h = 2$ las segundas diferencias, y así en adelante. Para la versión proporcional del método, las diferencias se miden en términos absolutos.

El método utiliza la minimización restringida, de una forma cuadrática, en relación con las diferencias entre las estimaciones desagregadas y una serie de indicadores. La función de penalización se puede especificar tanto como diferencias aritméticas como proporcionales. Este método es comúnmente utilizado por la facilidad y estabilidad de su implementación, así como por su robustez ante problemas de armonización. El problema se plantea como una minimización cuadrática de una función de penalización, minimizando los ajustes en el movimiento de la serie original, y expresado como:

$$min\ x_q \sum_{t=2}^{4y} [\ (X_t/X_{t-1}) - (I_t/I_{t-1})]^2 \qquad [7.9]$$

$$s.a. \sum_{4n-3}^{4n} X_t = A_n.\ n = 1, ..., y \qquad [7.10]$$

donde X_t es la serie trimestral a estimar, I_t es la serie trimestral disponible (indicador), A_n es la serie anual de la variable, $t = 1, ..., 4y$ es el índice de la serie trimestral y $n = 1, ..., y$ es el índice de la serie anual.

Este método no preserva explícitamente las tasas de variación trimestral de la serie indicadora, utilizadas comúnmente por los analistas de coyuntura económica. Para hacerlo, existe una alternativa no paramétrica desarrollada por Causey y Trager (1981), descrita posteriormente, que mantiene lo máximo

posible las tasas del indicador. Por último, el método Denton proporcional tiende a producir series más suavizadas y se considera una aproximación adecuada cuando la serie indicadora de referencia no muestra cambios inesperados entre periodos, o en los casos en los que la serie indicadora no es muy volátil.

Método Denton-Cholette

También conocido como Cholette-Dagum, es una generalización del método Denton que elimina los movimientos espurios al principio de las series de resultados. Cholette y Dagum (1994) proponen un procedimiento basado en una regresión por mínimos cuadrados generalizada, la cual toma en consideración la presencia de sesgo en el indicador de autocorrelación y heterocedasticidad en los errores de los datos originales, y se define según:

$$I_t = a_t + X_t + e_t, t = 1, ..., q \quad [7.11]$$

$$A_t = \sum_{4n-3}^{4n} X_t + w_n, n = 1, ..., y \quad [7.12]$$

donde X_t es la serie trimestral a ser estimada; I_t es la serie trimestral disponible (indicador); A_n es la serie anual de la variable; $t = 1, ..., 4y$ es el índice de la serie trimestral; $n = 1, ..., y$ es el índice de la serie anual; a_t es un efecto determinístico combinado; e_t es un error trimestral autocorrelacionado y heterocedástico; y w_n es el error heterocedástico de A_n incorrelacionado con e_t.

En una publicación posterior. Dagum y Cholette (2006a) definen un marco de regresión unificado construyendo una regresión de los indicadores de alta frecuencia sobre las restricciones de baja frecuencia, los efectos deterministas y los errores autocorrelacionados. El modelo es también adaptable a una forma multiplicativa y puede anidar otros métodos comunes de desagregación como en Chow-Lin, Fernández o Litterman.

Método de Chow y Lin

Este método desarrollado por Chow y Lin (1971), en adelante Chow-Lin, pertenece a la familia de los modelos basados en regresión, donde se incluyen también los de Fernández (1981) y Litterman (1983).

Los métodos autorregresivos se basan en modelos lineales generalizados sobre la serie anual y un conjunto de series indicadoras, en ellos se asume la existencia de una relación lineal entre las series de baja y alta frecuencia. Esta tipología

de aproximaciones difieren, esencialmente, en los modelos propuestos para la estructura de los residuos. Chow y Lin (1971), amplían el enfoque de mínimos cuadrados generalizados para la desagregación temporal, proponiendo una regresión univariante de los datos objetivo de baja frecuencia sobre indicadores de alta frecuencia. El método proporciona una solución óptima para la extrapolación, y propone un mecanismo para interpolar, armonizar y extrapolar series, basado en una regresión que usa, como variable explicativa, el indicador observado de alta frecuencia. Por tanto, se construye un modelo de regresión entre la variable no observada, de alta frecuencia (la serie armonizada), y una serie de indicadores relacionados de alta frecuencia observados, tal y como indica la expresión analítica siguiente:

$$X_t = \sum_{j=1}^{p} \beta_j I_{j,t} + u_t \qquad [7.13]$$

donde X_t es la serie trimestral a estimarse, I_t es la serie trimestral disponible (indicador), β_j es el coeficiente de regresión del indicador j anualizado y u_t es un error aleatorio $AR(1)$ con v_t. Como X_t no es observable, y por tanto desconocido, la expresión [7.13] no puede estimarse directamente, el método asume que la misma relación entre la variable y los indicadores en frecuencia trimestral se mantiene en la frecuencia anual, y propone una agregación del modelo a estimar según:

$$A_n = \sum_{j=1}^{p} \beta_j \; Ij, t + u_n^a \quad n \; = 1, ..., y \qquad [7.14]$$

donde A_n se define como una agregación de valores trimestral:

$$A_n = \sum_{4n-3}^{4n} X_t \qquad [7.15]$$

donde I_t es la serie trimestral del índice, β_j es el coeficiente de regresión del indicador j anualizado y u_n^a es un error anual $ARMA(1,1)$ derivado del autorregresivo $AR(1)$ de alta frecuencia. El método asume que los residuos de la serie de alta frecuencia siguen un proceso autorregresivo de orden 1, $AR(1)$, $u_t = \rho \; u_{t_1} + \epsilon_t$, donde ϵ es $WN(0, \sigma_\epsilon)$ [17].

Chow y Lin derivan un estimador BLUE [18] de los coeficientes β y de ρ. Más recientemente, Dagum y Cholette (2006a) demostraron que el modelo de Chow y Lin es un caso particular de su modelo de regresión aditivo, con una serie relacionada. El método asume un comportamiento $AR(1)$ para los residuos de la

[17] WN hace referencia a Ruido Blanco, del inglés *White Noise*.
[18] Mejor estimador lineal insesgado, o Best Linear Unbiased Estimator.

regresión, estimando su coeficiente autorregresivo con los datos, en lugar de mediante un calibrarlo por el usuario.

Método Causey y Trager

El método desarrollado por Causey y Trager (1981) plantea armonizar la serie de alta frecuencia minimizando los ajustes a la tasa de variación de la serie del índice, para el caso trimestral viene definida según las expresiones [7.16] y [7.17]. Como se observa en la expresión [7.16], la función objetivo es cuadrática y no lineal, y por tanto no es posible obtener una solución algebraica para la serie a estimar, las condiciones de primer orden del problema. La estimación de la serie de alta frecuencia armonizada debe realizarse a través de procedimientos de optimización no lineales.

$$min\ x_q \sum_{t=2}^{4y} [\ (X_t/X_{t-1}) - (I_t/I_{t-1})]^2 \qquad [7.16]$$

$$s.a. \sum_{4n-3}^{4n} X_t = A_n.\ n = 1, ..., y \qquad [7.17]$$

La eficiencia y robustez de los procedimientos utilizados en la estimación dependen de las particularidades de cada caso. Por lo tanto, esta aproximación puede dar lugar, en ocasiones, a problemas de falta de convergencia en la búsqueda de solución, estimaciones inexactas o resultados muy dependientes de los valores iniciales.

Parámetro autorregresivo y otros métodos

Los métodos restantes como Fernández o Litterman, se aplican a los casos en los que los indicadores de alta frecuencia y la serie anual no están cointegrados [19]. Fernández y Litterman asumen que los residuos trimestrales siguen un proceso no estacionario, es decir, $u_t = u_{t-1} + v_t$, donde v es un un modelo autorregresivo de orden 1, ($vt = \rho \cdot v_{t-1} + \epsilon_t$, donde ϵ Es $\mathcal{WN}(0, \sigma_\epsilon)$).

Fernández es un caso especial de Litterman, donde $\rho = 0$, y por tanto, u sigue una camino aleatorio aleatoria. La matriz de varianza-covarianza puede calcularse según como:

[19] La cointegración es una propiedad estadística de las variables en series temporales, donde al menos dos series temporales presentan una tendencia estocástica común. Esto indica que las series están vinculadas en el largo plazo y, aunque puedan desviarse temporalmente, retornarán a su relación de equilibrio a lo largo del tiempo (Enders, 2014).

$$\Sigma_L(\rho) = sigma_\epsilon^2[\Delta' H(\rho)' H(\rho)\Delta]^{-1} \quad [7.18]$$

donde Δ es la misma matriz de diferencias $n \times n$ que en el método Denton; $H(\rho)$ es una matriz $n \times n$ con 1 en su diagonal principal, $-\rho$ en su primera subdiagonal y 0 en el resto. Para el caso particular de Fernández, con el parámetro autoregresivo $\rho = 0$, la matriz de covarianzas resultado tendría la forma indicada en la siguiente ecuación:

$$\Sigma_L(0) = \sigma_\epsilon^2 \cdot (\Delta\Delta')^{-1} = \sigma_{epsilon}^2 \cdot \Sigma_D \quad [7.19]$$

En general, no existe un método único para el cálculo del parámetro autorregresivo ρ, el método Chow-Lin (Chow y Lin, 1971) propone un procedimiento iterativo, que infiere el parámetro de la autocorrelación observada de los residuos de baja frecuencia, u_l. En un enfoque diferente, Paige (1979) sugiere la maximización de la probabilidad de la regresión generalizada GLS. Por otro lado, Barbone *et al.* (1981) sugiere un minimizar la suma ponderada de los cuadrados de los residuos como indica la siguiente expresión:

$$RSS(\rho, \sigma_\epsilon^2, \beta) = u_l' \, (C \, \Sigma \, C')^{-1} \, u_l \quad [7.20]$$

Al contrario del enfoque de máxima verosimilitud, σ^2 no se cancela, por tanto, los resultados son sensibles a la especificación de σ, con diferentes implementaciones que conducen a estimaciones diferentes pero inconsistentes de ρ.

Más recientemente, se han desarrollado métodos alternativos a través de aproximaciones bayesianas Rojo-García y Sanz-Gómez (2005), cuyo objetivo es la estimación de la regularidad de las series en comparación con las técnicas clásicas que habitualmente muestran mayor volatilidad.

7.4.5 Estimadores de máxima verosimilitud

La estimación de máxima verosimilitud (MLE) es un enfoque fundamental en la estadística para estimar los parámetros de un modelo probabilístico. A pesar de su aparente simplicidad este método tiene propiedades teóricas sólidas, y ha demostrado su utilidad en una amplia variedad de aplicaciones prácticas. Este estimador permite trabajar con variables que no siguen una distribución normal, y se basa en la verosimilitud de que la muestra se haya generado con una función de distribución específica (Bickel y Doksum, 2015; Casella y Berger, 2021; DeGroot y Schervish, 2012).

La estimación de máxima verosimilitud fue introducido por Fisher (1922b), y ha sido ampliamente utilizado para estimar los parámetros de modelos probabilísticos. Se basa en un proceso que encuentra los valores de los parámetros de un modelo, que maximizan la función de verosimilitud que representa la probabilidad de observar los datos inferidos por el modelo.

Su estimación parte de una muestra aleatoria $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de observaciones independientes e idénticamente distribuidas (i.i.d.) de una distribución de probabilidad $f(x; \theta)$, donde θ es un vector de parámetros desconocidos. La función de verosimilitud, denotada por $L(\theta; \mathbf{x})$, se define como el producto de las funciones de densidad de probabilidad (PDF) de las observaciones individuales:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta) \tag{7.21}$$

El objetivo del método MLE es encontrar el valor de $\boldsymbol{\theta}$ que maximiza la función de verosimilitud:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta; \mathbf{x}) \tag{7.22}$$

Dado que el logaritmo es una función monótona creciente, maximizar la función de verosimilitud es equivalente a maximizar la log-verosimilitud, que se define como:

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \theta) \tag{7.23}$$

Para encontrar el valor de $\boldsymbol{\theta}$ que maximiza la log-verosimilitud, se toma la derivada parcial de $\ell(\boldsymbol{\theta}; \mathbf{x})$ con respecto a cada componente de $\boldsymbol{\theta}$ y se iguala a cero:

$$\frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, k \tag{7.24}$$

donde k es la dimensión de θ. Las ecuaciones resultantes se conocen como ecuaciones de puntuación, y su solución proporciona los estimadores de máxima verosimilitud de los parámetros desconocidos. En general, las ecuaciones de puntuación pueden ser no lineales y requerir métodos numéricos para encontrar su solución, como el método de Newton-Raphson o el algoritmo EM (Expectation-Maximization).

El método MLE tiene varias propiedades deseables. Bajo ciertas condiciones regulares[20], los estimadores de máxima verosimilitud son asintóticamente no

[20] Véase, por ejemplo, Casella y Berger (2021)

sesgados, eficientes y normalmente distribuidos:

$$\hat{\theta}_{\text{MLE}} \xrightarrow{d} \mathcal{N}(\theta, I(\theta)^{-1}), \tag{7.25}$$

donde $\boldsymbol{I}(\boldsymbol{\theta})$ es la matriz de información de Fisher, que se define como:

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_i \partial \theta_j}\right]. \tag{7.26}$$

La matriz de información de Fisher juega un papel importante en la inferencia estadística, ya que proporciona una medida de la incertidumbre en la estimación de los parámetros y permite construir intervalos de confianza y realizar pruebas de hipótesis.

Los estimadores de máxima verosimilitud han sido aplicados en diversos campos y a través de diferentes técnicas. Entre las técnicas más utilizadas se encuentra el algoritmo de maximización de la esperanza (Expectation-Maximization, EM), que fue propuesto por Dempster, Laird y Rubin (1977) y posteriormente perfeccionado por Meng y Rubin (1993). El algoritmo EM es especialmente útil en problemas de estimación donde los datos presentan cierto grado de falta de información o incompletitud, conocidos como datos ausentes o datos censurados.

El algoritmo EM es un método iterativo que busca maximizar la función de verosimilitud incompleta mediante la actualización alternativa de dos pasos: el paso de esperanza (E-step) y el paso de maximización (M-step). En el paso de esperanza, se calcula la esperanza condicional de la verosimilitud completa, dada la muestra observada y los valores actuales de los parámetros. En el paso de maximización, se actualizan los parámetros maximizando la esperanza condicional calculada en el paso de esperanza. El algoritmo EM se repite hasta que se alcanza la convergencia en los parámetros estimados.

Capítulo 8

Índice de precio de alquiler

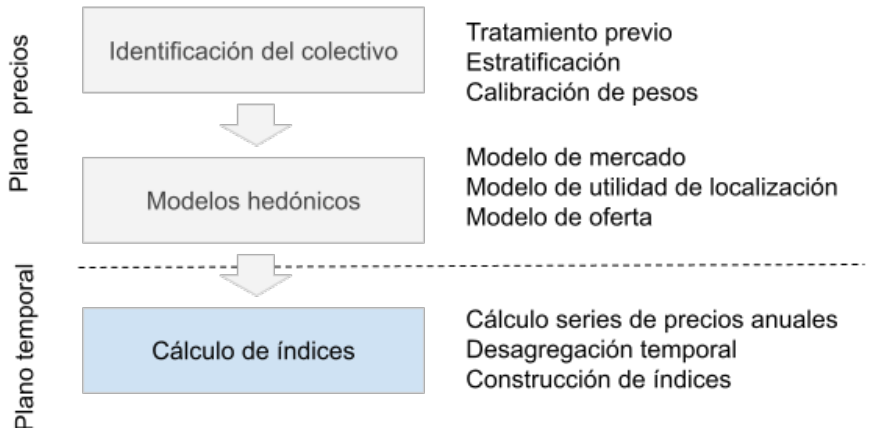
"Revisa todos los detalles que afectan a tu trabajo."

— Arthur C. Nielsen

8.1 Introducción

Este capítulo finaliza la metodología con la construcción de los índices de precios los valores de mercado y de oferta. Se apoya en los resultados de los capítulos anteriores, resumidos en la Figura 8.1, y que parten de la estratificación y cálculo de los pesos poblacionales, en el Capítulo 3; para, posteriormente, elaborar los modelos hedónicos de mercado y de oferta, que se presentan en los Capítulos 4, 5 y 6. Dichos resultados permiten la construcción de las series temporales de precios anuales y mensuales del Capítulo 7; y que finalmente, en el presente capítulo, sirven para la elaboración de los índices de los precios definitivos.

Figura 8.1. Resumen del proceso general de la metodología



Fuente: elaboración propia.

El planteamiento propuesto en esta Tesis se basa en las recomendaciones del manual para la construcción de índices del precio de la vivienda residencial de Eurostat (2014). En dicho manual, se insiste en la armonización de los resultados entre mercados, la gestión de la diversidad de características en las viviendas, la estandarización de los métodos hedónicos, y que los resultados se generen con un alto nivel de frecuencia y de la forma más actualizada posible. Paralelamente, se intenta mantener los criterios metodológicos del Índice del Precio de la Vivienda del INE (2016a), que son más específicos a la realidad inmobiliaria española pero que incorporan las recomendaciones europeas mencionadas anteriormente.

La metodología de construcción de los índices sigue tres pasos: el primero, crea los índices básicos para los estratos en los que se divide la población; el segundo, calcula las ponderaciones para construir los agregados de los índices; y tercero, construye los índices de precios para alquiler y oferta, para un conjunto de agregados de tipo funcional y zonal.

El contenido del presente capítulo se estructura en dos partes: la primera, describe el enfoque metodológico seguido para la construcción del índice, ahondando el criterio de estratificación y el tipo de índice utilizado; y la segunda, presenta los resultados obtenidos, con un estudio de la calidad de las estimaciones, desde el punto de vista de estructura y capacidad para anticipar eventos futuros. Para los resultados obtenidos, se realiza una interpretación en detalle de los valores desglosados por criterios funcional y zonal, y contrastando la metodología con el índice de precios del alquiler experimental elaborado por el INE (2021b), así como su capacidad con predictor de movimientos futuros en el mercado de la vivienda.

8.2 Metodología

El proceso del cálculo del indicador recoge dos aspectos fundamentales del mercado de alquiler de la vivienda: los precios de las rentas y la composición del stock de viviendas alquiladas. Ambos aspectos se combinan en un índice de precios, que es de tipo Fisher[1] encadenado, de características muy similares a los usados en el cálculo del IPV o del IPC/IPCA del INE[2] .

La decisión de utilizar un índice de Fisher se fundamenta en la recomendación de Eurostat (2014) de aplicar preferentemente un índice de tipo superlativo, por varios motivos: primero, cuenta con un mayor número de propiedades económicas y axiomáticas de número índice (Balk, 1995; Diewert, 1976); y segundo, que es

[1] En la literatura es habitual aplicar cualquiera de estos tres tipos de índices: Laspeyres, Paasche o Fisher (Griliches, 1990).

[2] En realidad tanto el IPC como el IPC/IPCA se calculan como índices de Laspeyres encadenados.

menos propenso a sesgos (2020), como, por ejemplo, el de sustitución[3].

El control de los cambios de composición de la calidad de los bienes a lo largo del tiempo, se realiza a través de los modelos hedónicos y el uso de índices encadenados. En nuestro caso, es una cuestión clave a controlar dada la heterogeneidad natural del mercado de la vivienda, y porque, además, la muestra original procede de los precios de oferta de un portal de Internet.

8.2.1 Criterio de estratificación

La estratificación usa dos criterios de desagregación: el funcional y el zonal, combinables si el estrato funcional está presente en la zona de estudio. En el caso de zonal, se utiliza la misma división aplicada en la calibración[4], definida como:

- Desglose macro-geográfico entre zonas la ciudad de Madrid y zonas del resto de la Comunidad de Madrid.
- Para el caso de Madrid, se usa una agrupación basada en los barrios, con un total de 112 zonas (de un total de 128 barrios originales).
- En el resto de la Comunidad, se trabaja en regiones municipales o agregaciones de municipios. Estas agregaciones se basan en la unión de zonas con insuficiente soporte muestral, que sean cercanas y con características inmobiliarias similares.

Tabla 8.1. Variables usadas para la estratificación funcional

Variable	Descripción	Valores
SUT	Superficie útil	4 niveles
NHAB	Número de habitaciones	4 niveles
ANOCONSC	Año de construcción	4 niveles
TIPOEDIF	Tipo de edificio	4 niveles
TIPOCASA	Tipo de vivienda	3 niveles
ZONARES	Tipo de zona residencial	3 niveles: renta baja, media y alta
INTERINPSP	Intervalo de ingresos mensuales netos totales	3 niveles: ingresos bajos, medios y altos
FACTORGASTOT6	Gastos familiares	3 niveles, gasto bajo-medio, medio alto, alto
DENSI	Densidad de población del municipio	3 niveles
TAMAU	Tamaño del municipio en población	5 niveles

Fuente: elaboración propia

[3]El sesgo de sustitución se refiere a la distorsión inducida en un índice de precios por la aparición los productos sustitutivos, en la vivienda se puede referir a la sustitución de vivienda en propiedad por la alquiler, cambio del precio relativo entre dormitorios y baños o cambio de preferencias de una zona a otra. Esta cuestión se desarrolla con mayor profundidad en Tripplet (1996) y Hill (2006)

[4]Para más detalle ver el epígrafe 3.2.2.2 del Capítulo 3.

El desglose funcional se realiza sobre 11 criterios de tipo zonal, de características físicas del inmueble y condiciones socioeconómicas del hogar, los criterios funcionales[5] se muestran en detalle en la Tabla 8.1.

Las ponderaciones recogidas por los elevadores muestrales permiten componer la población del colectivo sobre una muestra representativa, donde el elevador muestral indica el peso de dicha representación. La importancia de cada estrato muestral se calcula mediante la proporción del gasto total en dicho estrato en relación con el gasto en vivienda.

Los pesos se consideran una variable de flujo que varía año tras año, aún con la salvedad de que en el mercado del alquiler existan contratos plurianuales. Esta aproximación garantiza una notable estabilidad temporal a la medida, más si se compara con su equivalente del mercado de compraventa (transacciones registradas por los notarios en un periodo).

La estructura de las ponderaciones se obtiene con información referida a los dos últimos años disponibles, tal y como recomienda Eurostat (2014), de forma se logran dos objetivos:

- Asegurar la estabilidad de la estructura de las ponderaciones. Ya que este método, basado en medias móviles, suaviza posibles cambios de calidad y composición bruscos.
- Mejorar la representatividad del índice. Pues cuantos más años intervienen en el cálculo, más estratos y tipos de vivienda estarían representados, y, por lo tanto, mejor es el ajuste por cambios de calidad y composición. En caso contrario, un intervalo más amplio no sería deseable, ya que el indicador arrojaría una visión distorsionada de la situación actualizada del alquiler.

Debido al dinamismo de este mercado mercado, se actualiza anualmente la estructura de los pesos, para que se reflejen, de forma fiel y real, la coyuntura inmobiliaria mediante un índice encadenado.

Para calcular las ponderaciones se toman dos valores para cada estrato e: el primero, es la superficie útil media en m² para el estrato[6], y el segundo, los precios medios por superficie útil. Ambas magnitudes se corresponden a las cantidades (Q) y los precios (P) utilizados en el proceso. Los pesos W_e^a, para el estrato e y el año a, se expresan como un índice de tipo Fisher, calculado como la media geométrica de dos índices encadenados, uno de Paasche y otro de Laspeyres. Su forma funcional es la siguiente:

[5] Para más información sobre el significado y mayor información cada una de las variables, véase el apartado 3.2.2.1 del Capítulo 3.

[6] Que se calculan como la media ponderada de la superficie útil de las viviendas, usando como pesos los factores de elevación.

$$W_e^a = \frac{\sqrt{Q_e^{(a-1,a-2)} \times \hat{P}_e^{a-1} \times \hat{P}_e^{(a-1,a-2)} \times Q_e^{a-1}}}{\sum_{\forall e} \sqrt{Q_e^{(a-1,a-2)} \times \hat{P}_e^{a-1} \times \hat{P}_e^{(a-1,a-2)} \times Q_e^{a-1}}} \quad [8.1]$$

donde $Q_e^{(a-1,a-2)}$ es la cantidad media, en metros cuadrados de superficie útil, en el estrato e para los dos años anteriores a a. Mientras que, $\hat{P}_e^{a-1}$ es el precio medio en €/m² útiles del estrato en el año anterior. De forma análoga, $\hat{P}_e^{(a-1,a-2)}$ es la media del precio de los dos años anteriores para e, y Q_e^{a-1} los metros útiles medios del año anterior.

El índice es de tipo hedónico de imputación doble, utilizando las cantidades y los precios calculados por los modelos hedónicos, tanto en el año base como en siguientes periodos.

La ponderación de cualquier agregado A, ya sea del tipo funcional o geográfico, se obtiene como la suma de las ponderaciones de los estratos que comprende el agregado:

$$W_A^a = \sum_{e \in A} W_e^a \quad [8.2]$$

donde W_A^a es el peso del agregado A para el año a, y W_e^a el peso de cada estrato e.

8.2.2 Cálculo de índices

Como se indicó en el epígrafe previo, la fórmula general utilizada para el cálculo del índice es un índice de Fisher encadenado con actualización anual de las ponderaciones, que a su vez utiliza modelos hedónicos anuales para el control por calidad y características de los inmuebles. Esto permite aplicar futuras revisiones metodológicas, tanto en el desarrollo de los modelos, como en la estratificación (añadir nuevos estratos, por ejemplo). Además, el método encadenado permite ir acomodando las nuevas condiciones en el mercado, bien en los precios o en la composición muestral.

Existen tres periodos de referencia para el índice anual propuesto:

- Periodo de referencia del índice: es momento temporal base (año 2011) para todos los índices, se le asigna un valor de 100.
- Periodo de referencia de las ponderaciones: es al que se refieren los datos utilizados en el cálculo de las ponderaciones. Cada año se actualizan las ponderaciones del índice con la información disponible de la calibración y los precios estimados del alquiler, de los dos años anteriores. Para los años 2011

y 2012, al no contar con información de los dos años anteriores completos, se estiman los pesos usando los datos del año 2011.

- Periodo de referencia de los precios: es aquel con cuyos precios se comparan con los precios corrientes, en este caso sería el precio del año anterior: $a-1$.

Los índices elementales, o agregados elementales, son los índices con el menor nivel de agregación y para los que no se aplican las ponderaciones. Representan al estrato básico que recoge un tipo de vivienda, por desagregación funcional o zonal.

El índice elemental I_e^a, se calcula como el cociente entre el precio estimado del modelo para las viviendas, perteneciente al periodo actual, y el precio estimado para el año anterior.

$$I_e^a = \frac{\hat{P}_e^a}{\hat{P}_e^{a-1}} \times 100 \qquad [8.3]$$

donde, $\hat{P}_e^a$ es el precio estimado para el estrato e y el año a, y $\hat{P}_e^{a-1}$ el precio estimado para el estrato e y el año $a-1$.

Los índices agregados, dada cualquier agregación A, se calculan como la suma ponderada de la contribución de sus distintos estratos, mediante:

$$I_A^a = \sum_{e \in A} W_e^a \times I_e^a \qquad [8.4]$$

donde I_A^a es el agregado para el año a, I_e^a el índice para el estrato e y W_e^a su correspondiente peso.

Los índices finales, con base 2011, se calculan multiplicando la serie de índices encadenados, según:

$${}_{2011}I_A^a = 100 \times \prod_{y=2011}^{a} \frac{I_A^a}{100} \qquad [8.5]$$

O en su forma más compacta:

$${}_{2011}I_A^a = {}_{2011}I_A^{a-1} \times \frac{I_A^a}{100} \qquad [8.6]$$

Por ejemplo, el índice correspondiente a 2014 para una agregación A, se calcularía en función de los índices encadenados como:

$${}_{2011}I_A^{2014} = \frac{I_A^{2011}}{100} \times \frac{I_A^{2012}}{100} \times \frac{I_A^{2013}}{100} \times \frac{I_A^{2014}}{100} \times 100 \qquad [8.7]$$

8.3 Resultados

A continuación, se evalúan los índices anuales de precios finales de alquiler y de oferta, según los desgloses funcional y geográfico. El análisis comprende el estudio de varios aspectos de tipo cualitativo y cuantitativo, abundando en la consistencia de los resultados y su potencial de aplicación, que se concretan en los siguientes puntos y se desarrollan en los próximos subepígrafes:

1) Evaluación de la calidad de las series desde un punto de vista de estructural.
2) Comparativa consistencia de los índices alquiler y oferta.
3) Análisis de las series generadas, tanto de alquiler como de oferta, desde la perspectivas funcional y geográfica.
4) Comparativa con el IPVA publicado por el INE (2021b).
5) Capacidad predictiva de los índices.

Todos los datos y gráficas presentadas pueden consultarse en los enlaces recogidos en el Anexo I.

8.3.1 Evaluación de la calidad de las series

Se han analizado un total de 244 series de cada uno de los dos índices construidos: precio del alquiler y precio de oferta. Dichas series proceden de la estratificación en 11 categorías funcionales y una zonal, definidas en el epígrafe 8.2.1, que se distribuyen en 177 estratos zonales y 67 funcionales.

La calidad estructural de las series se evalúa en función del comportamiento de las diferencias interanuales del índice. Entre las métricas más comúnmente utilizadas para este fin se encuentran: el coeficiente de variación, la desviación estándar, la autocorrelación, y las pruebas de bondad de ajuste y de estabilidad temporal. Estas medidas permiten analizar la variabilidad, dispersión y precisión de las estimaciones del índice y, en consecuencia, evaluar su calidad (Eurostat, 2013).

Tabla 8.2. Autocorrelación de las diferencias anuales

	Autocorrelación		
Índice	$\rho(\Delta I_{t,t-1}, \Delta I_{t-1,t-2})$	$\rho(\Delta I_{t-1,t-2}, \Delta I_{t-2,t-3})$	$\rho(\Delta I_{t-2,t-3}, \Delta I_{t-3,t-4})$
Alquiler	0,85	0,78	0,66
Oferta	0,86	0,73	0,53

Fuente: elaboración propia

La Tabla 8.2 muestra la autocorrelación (ρ) de las diferencias interanuales de los valores de los índices I, observándose la particular fuerza y consistencia temporal

de esta relación para la serie de alquiler.

En términos de variabilidad, presenta un comportamiento más estable en las series de alquiler, calculada como el Coeficiente de Variación de Pearson[7] (CoV), calculado mediante la expresión:

$$CoV^e = \frac{\sigma(\Delta I^e)}{|\mu(\Delta I^e)|} \quad [8.8]$$

donde CoV^e es el coeficiente de variación para el estrato e, $\mu(\Delta I^e)$ la media de las diferencias anuales, y $\sigma(\Delta I^e)$ su desviación típica.

El valor de la medida anterior para las dos series temporales calculadas se recoge en la Tabla 8.3. En ella se observa que aunque el índice de alquiler ofrece variaciones interanuales relativamente extremas (-20,4 mínima y 50,0 máxima), el resto de la distribución mantiene un comportamiento moderado. Esto se evidencia a través del valor del percentil 99 de la serie, que asciende a 10,6.

Tabla 8.3. Comparativa descriptivos estructurales en indices de alquiler y oferta

	Índices			Diferencias anuales índices						
Variable	**Media**	**Stdev.**	**CoV**	**Media**	**Stdev.**	**CoV**	**Min**	**p01**	**Max**	**p99**
Oferta	101,7	12,2	12,0%	6,4	2,8	6,3%	-19,9	-8,8	28,2	18,4
Alquiler	93,9	6,8	7,2%	5,2	0,3	5,5%	-20,4	-10,5	50,0	10,6

Fuente: elaboración propia

Tal y como sucedía con los modelos de valoración hedónica, el origien de buena parte de las anomalías y alta volatilidad se origina en aquellos estratos que tienen una muestra más pequeña o irregular.

La hipótesis anterior se puede corroborar con los valores observados en la Tabla 8.4, que recoge una serie de estratos con un alto grado de CoV (valores superiores al 100%). En todos estos casos, las series más volátiles se concentran en los estratos menos numerosos o con inmuebles singulares: aquellos con mayor superficie (por ejemplo, Chalé o 5 o más habitaciones), viviendas unifamiliares (como, por ejemplo, Arganda del Rey o Rivas-Vaciamadrid) o áreas rurales[8] (por ejemplo, Colmenar de Oreja- Chinchón).

[7] El Coeficiente de Variación de Pearson, también conocido como índice de variabilidad relativa, es un instrumento estadístico que proporciona una medida normalizada de la dispersión de una distribución de datos. Se calcula dividiendo la desviación estándar por el valor absoluto de dicha media. Esto permite obtener una razón que indica el grado de heterogeneidad de los valores de la variable aleatoria en relación con su promedio, facilitando así la comparación de dicho grado de dispersión en distintas poblaciones.

[8] Las zonas rurales se caracterizan por un número bajo e inestable de inmuebles en la muestra.

Tabla 8.4. Estratos con mayor variabilidad en el índice de alquiler

Ámbito	Factor	Nivel	Rural	Media dif.	CoV dif.
Resto	CODIGO	Arroyomolinos	Sí	3,8	147,9%
Resto	NHABIT	5 o más habitaciones		13,5	139,1%
Madrid	TIPOCASA	Chalé o casa grande		16,7	139,0%
Madrid	TIPOEDIF	Vivienda unifamiliar adosada o pareada		16,7	139,0%
Resto	CODIGO	Área de Colmenar de Oreja - Chinchón	Sí	8,0	136,4%
Resto	CODIGO	Arganda del Rey	No	6,7	130,7%
Resto	CODIGO	Rivas-Vaciamadrid	No	5,6	126,9%
Madrid	ZONARES	Urbana de lujo		4,8	126,4%
Madrid	NHABIT	5 o más habitaciones		10,4	126,0%
Resto	ZONARES	Urbana de lujo		8,9	118,4%

Fuente: elaboración propia

Debido a que la tendencia general del mercado (series más agregadas) comienza con un descenso hasta 2014 seguido por un periodo de tendencia ascendente, se espera que las series desagregadas mantengan un comportamiento similar (al menos en términos de cambios en la tendencia). Para evaluar las discrepancias con el escenario general, se mide número de cambios de signo (Tabla 8.5), observándose como el número de cambios predominante es 1, mientras que una frecuencia de 4 cambios es minoritaria. Madrid ofrece un mayor porcentaje de series con un solo cambio y, también, el índice de alquiler es más estable según este criterio.

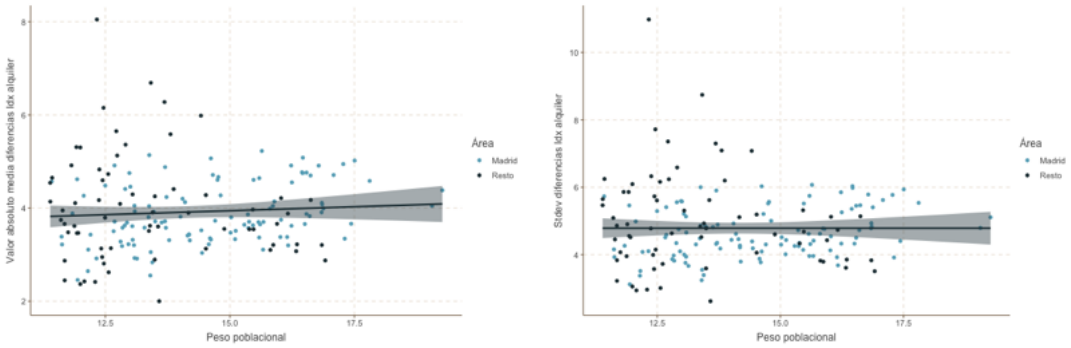
Tabla 8.5. Distribución del número de cambios de signo en porcentaje

		Número de cambios					
Índice	**Ámbito**	**1**	**2**	**3**	**4**	**5**	**6**
Oferta	Madrid	85,9%	9,9%	3,5%	0,7%		
	Resto	62,7%	17,6%	13,7%	3,9%	2,0%	
	Todas	76,2%	13,1%	7,8%	2,0%	0,8%	
Alquiler	Madrid	69,7%	16,2%	12,7%	1,4%		
	Resto	58,8%	10,8%	22,5%	2,9%	2,9%	2,0%
	Todas	65,2%	13,9%	16,8%	2,0%	1,2%	0,8%

Fuente: elaboración propia

Las medias de las desviaciones de las diferencias anuales muestran que no existe una relación entre el tamaño de la zona y las tasas de variación de los precios, ni tampoco hay diferencias en este sentido en función de si la vivienda está en Madrid o fuera (Figura 8.2).

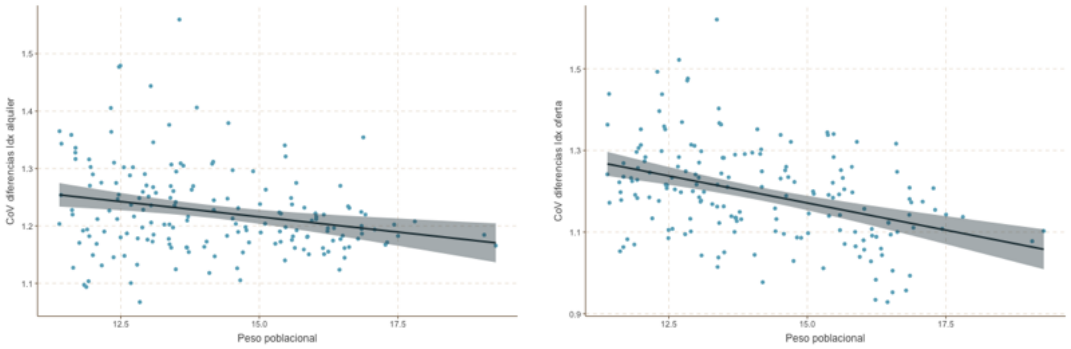
Figura 8.2. Medias de las diferencias en los índices



Fuente: elaboración propia.

Como se indicaba en el epígrafe 3.4.2, la variabilidad de los precios es mayor cuando más pequeñas son las zonas, por tanto, es esperable que suceda lo mismo en el índice final. Para ello, se analiza la relación entre el tamaño de cada estrato en función de su peso poblacional y el nivel de variabilidad del índice, calculado como coeficiente de variación de Pearson. Se comprueba gráficamente, en la Figura 8.3, como es así de forma intensa para la serie de oferta y que para el alquiler.

Figura 8.3. Coeficiente de variación sobre las diferencias de los índices de alquiler y oferta



Fuente: elaboración propia.

La distribución de los dos índices, agrupadas por quintiles por peso poblacional, para los estratos de zona geográfica (véase Tabla 8.6), confirman que el CoV decrece con el tamaño de la muestra (aunque de forma más irregular en el de oferta). Sin embargo, también se aprecia que el valor medio y la desviación absoluta crecen cuanto mayor es la muestra.

Tabla 8.6. Series de índices con mayor variabilidad por zona geográfica

		Índices		Diferencias anuales					
Índice	**Quintil**	**Media**	**Stdev.**	**Media**	**P20**	**P40**	**P60**	**P80**	**CoV**
Oferta	1	97,9	6,5	5,0	1,6	3,1	4,9	8,1	120,2%
	2	96,7	8,2	4,8	1,9	3,4	4,9	7,5	123,8%
	3	96,8	9,2	4,7	1,5	3,2	4,9	7,2	125,7%
	4	108,0	14,2	7,9	3,2	5,6	9,0	12,5	94,4%
	5	113,3	15,4	6,2	2,3	4,4	6,8	9,4	98,4%
Alquiler	1	94,7	6,3	3,4	1,4	2,6	3,3	4,8	123,6%
	2	92,4	5,7	3,3	1,2	2,4	3,4	4,8	117,7%
	3	91,1	5,7	3,7	1,8	3,0	3,9	5,5	113,6%
	4	94,0	7,1	4,3	2,3	3,2	4,5	6,3	108,8%
	5	96,7	7,9	5,2	3,1	4,4	5,8	7,3	86,9%

Fuente: elaboración propia

Por otra parte, la consistencia temporal de las cantidades (Q) utilizadas para los índices del alquiler se analiza a través de las diferencias de la distribución multidimensional que representan para un cada periodo. Para comparar estas diferencias se utiliza la distancia Hellinger, introducida en el epígrafe 3.4.2, que en este caso se basa en las diferencias en probabilidad para cada estrato entre dos años distintos. Esta medida, por consiguiente, sintetiza si existen cambios relevantes en la estructura de cantidades entre dos conjuntos (con múltiples dimensiones).

La Tabla 8.7. muestra los resultados de comparar las distancias entre cada par de años posibles. Se observa como la distribución se mantiene similar a lo largo del tiempo, por lo que se concluye la estabilidad de esta dimensión de los índices. En términos absolutos, las diferencias Hellinger son de media $4,8 \times 10^{-5}$, la que indica que la distribución conjunta es prácticamente estable en todos los años, con ligeros cambios a lo largo del tiempo. El caso que muestra mayores diferencias es el 2011 con los periodos entre 2017 y 2019, siendo en cualquier caso unos valores mínimos en distancia.

Tabla 8.7. Distancia Hellinger cantidades del índice entre pares de años (A y B), escalados a tanto por 1000

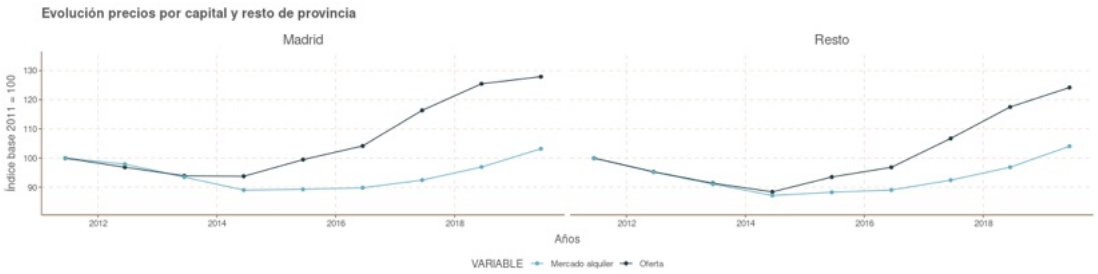
Año A	Año B							
Año	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
2011	0,06	0,07	0,08	0,07	0,07	0,10	0,13	0,14
2012		0,01	0,02	0,01	0,02	0,05	0,07	0,08
2013			0,01	0,01	0,01	0,03	0,06	0,07
2014				0,01	0,01	0,03	0,05	0,06
2015					0,01	0,04	0,07	0,07
2016						0,04	0,06	0,07
2017							0,03	0,04
2018								0,01

Fuente: elaboración propia

8.3.2 Consistencia entre índices de alquiler y oferta

Desde un punto de vista económico, el stock en oferta, en mercados activos, se incorpora progresivamente como inmuebles alquilados, de forma que los inmuebles rotan del mercado del alquiler al de oferta y viceversa. En este sentido, Kokot (2015), comprueba experimentalmente que las series de precios de oferta y mercado están cointegradas y muestran a la oferta como un indicador adelantado en 5 meses de los precios reales. En nuestro caso, en el Capítulo 6, se demuestra como la tendencia de la serie de mercado tiene mayor correlación con el valor del oferta del año anterior, que con el del año corriente.

Figura 8.4. Comparativa evolución del precio de oferta y de alquiler, desglose por capital y resto de provincia



Fuente: elaboración propia.

Para comprobar la consistencia de mercado del índice, se ha evaluado si el indicador de mercado materializa los cambios del índice de oferta. Las series de oferta y mercado, recogidas en la Figura 8.4, indican que la serie de oferta está adelantada en torno a un periodo con respecto a la de alquiler, lo que es concordante con lo visto en el epígrafe anterior, en la Tabla 8.2, al comparar las

diferencias.

Numéricamente, en Madrid el coeficiente de correlación de Pearson del índice de mercado y la serie de oferta del año corriente $\rho(I_t^a, I_{t-1})$ es de 0,65, mientras que con el índice de oferta del año anterior $\rho(I_t^o, I_{t-1})$ es del 0,90. Para el resto de la Comunidad las proporciones cambian ligeramente, con un valor para las series del mismo año es de 0,90, mientras que, para el índice de oferta del año anterior es del 0,87.

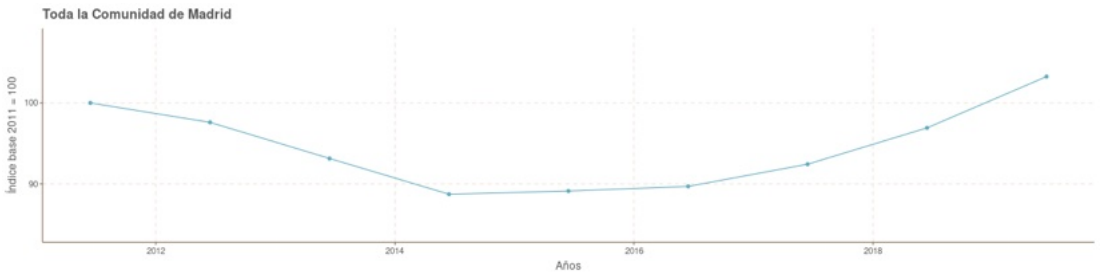
8.3.3 Análisis funcional y geográfico

A continuación, se examinan las series del índice de precios de acuerdo con los desgloses funcional y geográfico. Este enfoque multifacético tiene como objetivo ofrecer una visión integral y detallada de la evolución de los precios en los diferentes submercados de viviendas en la Comunidad de Madrid. Paralelamente, y con el fin de complementar la perspectiva anterior, se lleva a cabo el análisis tanto para los datos de alquiler como para los de oferta.

8.3.3.1 Análisis de las series de alquiler (desglose funcional)

El índice general de precios, Figura 8.5, muestra un comportamiento decreciente del alquiler desde el año base hasta el año 2016[9], cuando se produce el cambio de tendencia al crecimiento sostenido hasta 2019.

Figura 8.5. Índice de precio del alquiler residencial general



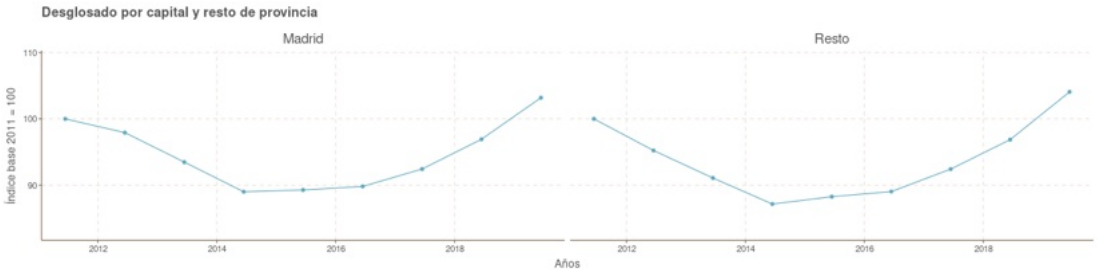
Fuente: elaboración propia.

El desglose por capital y resto de provincia muestra comportamientos diferentes, como se observa en la Figura 8.6. La capital cae desde el año base hasta el año 2016, con un ligero repunte hasta el año 2017, acelerándose a partir de entonces. El resto de la provincia muestra un comportamiento similar, con una caída sostenida hasta el año 2014, una estabilización de precios hasta el año 2017

[9] Para representar las series, los valores anuales se muestran a mitad de periodo, el 15 de junio de su correspondiente año.

y con un ligero crecimiento posterior, que comienza a ser significativo a partir del año 2018.

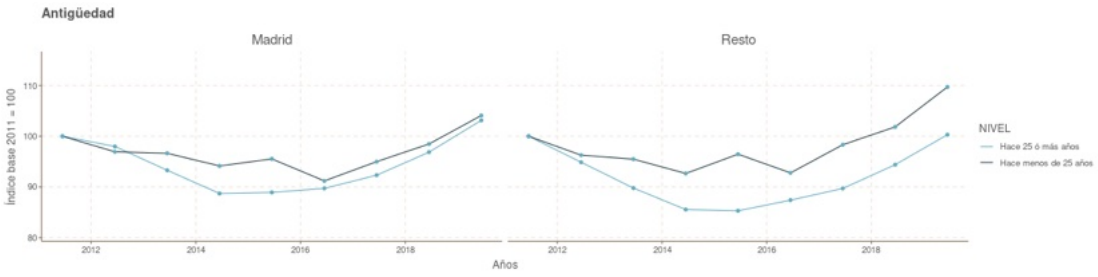
Figura 8.6. Índice de precio del alquiler desglosado por Madrid y resto de provincia



Fuente: elaboración propia.

Se ha construido el índice desglosado en función de las distintas características de la vivienda. Si atendemos al desglose alguna de las características constructivas, como por ejemplo, el año de construcción, las viviendas más nuevas muestran un comportamiento más resistente a la crisis que las viviendas antiguas, como vemos en la Figura 8.7. Sin embargo, estas últimas, en la fase de recuperación, muestran crecimientos similares. Es importante destacar que para las viviendas nuevas, se aprecia un ligero cambio de tendencia en 2015 que no atiende a condiciones de mercado, y podría atribuirse a efectos de composición.

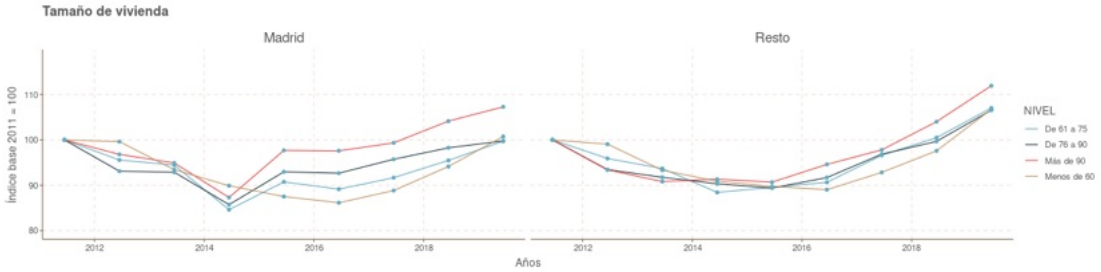
Figura 8.7. Índice de precio del alquiler desglosado por año de construcción



Fuente: elaboración propia.

En cuanto al criterio de segmentación por superficie útil, se observa un mejor comportamiento en las viviendas de mayor tamaño. En contraste, las viviendas más pequeñas experimentaron caídas más pronunciadas durante el periodo 2011-2016, como se ilustra en la Figura 8.8.

Figura 8.8. Índice de precio del alquiler desglosado por superficie útil



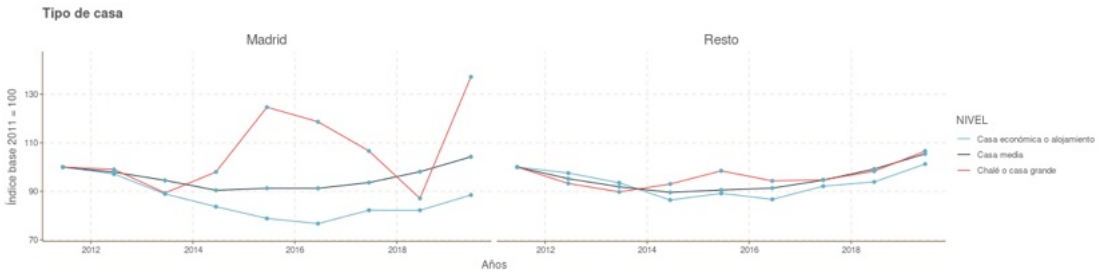
Fuente: elaboración propia.

El desglose en función del tipo de vivienda muestra que las viviendas de tipo unifamiliar tuvieron un comportamiento más estable en el tiempo, con un mayor incremento de precios en el resto de provincia (Figura 8.9).

En general, las viviendas más económicas son las que ofrecen el peor comportamiento, tanto en la capital como en el resto de zonas, mientras que el tipo medio cuenta con menores caídas en precios y crecimientos más intensos.

Por otra parte, se observa un comportamiento muy irregular del índice de precios de las viviendas unifamiliares de la capital, que se puede atribuir a dos factores: el primero, que este tipo de inmuebles es poco común en la capital y, cuando existe, es un producto bastante exclusivo, con un número importante de variables omitidas, sobre el que es más complicado generalizar la valoración; el segundo, debido a una muestra muy pequeña que ofrece medidas más irregulares.

Figura 8.9. Índice de precio del alquiler desglosado por tipo de vivienda



Fuente: elaboración propia.

Se resume el desglose de los índices, desagregados funcionalmente por criterios constructivos, en la Tabla 8.8.

Tabla 8.8. Índices de precios desglose funcional por estructura

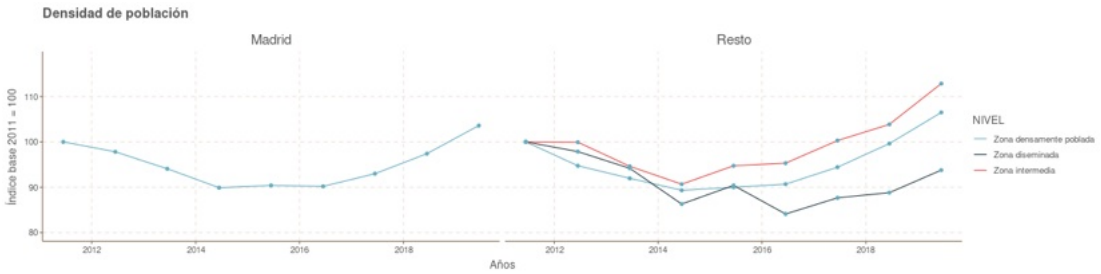
Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Antigüedad	Hace 25 ó más años	Madrid	100	98.0	93.3	88.7	88.9	89.7	92.3	96.9	103.1
		Resto	100	94.9	89.8	85.6	85.3	87.4	89.7	94.4	100.3
	Hace menos de 25 años	Madrid	100	96.9	96.6	94.2	95.5	91.2	95.0	98.5	104.1
		Resto	100	96.3	95.5	92.7	96.5	92.8	98.3	101.8	109.7
Número de habitaciones	1 o 2 habitaciones	Madrid	100	98.0	92.8	87.9	88.0	86.8	89.6	94.4	99.6
		Resto	100	95.4	91.8	87.8	88.8	87.8	92.4	95.9	102.2
	3 habitaciones	Madrid	100	95.4	92.8	86.7	90.8	91.2	92.8	97.4	103.0
		Resto	100	95.1	91.2	88.1	87.2	90.4	93.5	99.1	106.7
	4 habitaciones	Madrid	100	99.0	95.6	88.5	94.1	95.8	94.5	100.8	108.1
		Resto	100	95.3	92.1	91.2	88.5	91.6	94.9	100.1	107.4
	5 o más habitaciones	Madrid	100	101.5	89.7	84.0	98.1	101.0	86.5	94.9	119.1
		Resto	100	90.0	69.6	71.5	91.3	76.5	77.8	74.7	111.3
Tamaño de vivienda	De 61 a 75	Madrid	100	95.6	94.4	84.6	90.7	89.1	91.7	95.5	99.7
		Resto	100	95.9	93.7	88.4	89.5	90.7	96.6	100.5	107.0
	De 76 a 90	Madrid	100	93.1	92.9	85.7	93.0	92.7	95.7	98.3	99.8
		Resto	100	93.4	91.7	90.3	89.4	91.7	96.9	99.7	106.5
	Menos de 60	Madrid	100	99.6	93.4	89.9	87.5	86.2	88.8	94.1	100.7
		Resto	100	99.1	93.3	90.9	89.7	89.0	92.8	97.6	106.8
	Más de 90	Madrid	100	96.8	94.9	87.3	97.7	97.6	99.3	104.1	107.3
		Resto	100	93.4	90.8	91.4	90.7	94.6	97.8	104.0	111.9
Tipo de casa	Casa económica o alojamiento	Madrid	100	97.1	88.9	83.7	78.8	76.7	82.2	82.2	88.5
		Resto	100	97.5	93.5	86.4	89.2	86.7	92.1	93.9	101.2
	Casa media	Madrid	100	97.9	94.5	90.4	91.3	91.3	93.6	98.1	104.2
		Resto	100	95.2	91.9	89.6	90.5	91.3	94.6	99.2	105.4
	Chalé o casa grande	Madrid	100	99.0	89.3	98.0	124.7	118.6	106.6	87.1	137.1
		Resto	100	93.2	89.8	93.0	98.4	94.3	94.8	98.3	106.6
Tipo de edificio	Con 10 ó más viviendas	Madrid	100	97.5	93.8	89.6	90.3	90.3	92.7	97.2	103.2
		Resto	100	94.9	91.3	89.0	90.1	90.3	94.1	99.3	104.2
	Con menos de 10 viviendas	Madrid	100	101.1	97.2	92.2	89.1	88.5	91.0	93.5	101.6
		Resto	100	97.4	94.2	88.2	89.9	90.3	93.2	94.9	104.9
	Vivienda unifamiliar adosada o pareada	Madrid	100	99.0	89.3	98.0	124.7	118.6	106.6	87.1	137.1
		Resto	100	94.4	83.4	89.4	91.7	85.6	84.6	87.7	100.1
	Vivienda unifamiliar independiente	Resto	100	91.5	90.5	89.9	97.0	95.1	95.0	97.2	104.2

Fuente: elaboración propia

Cuando el análisis se realiza por características de zona (Figura 8.10), se

observa que aquellas que están diseminadas (rurales) ofrecen los descensos más acusados y tasas de crecimiento menores, comparados con las zonas intermedias y las densamente pobladas. Estas diferencias son atribuibles a que las zonas urbanas tienen un mayor nivel de demanda que las rurales y, por tanto, la presión de la demanda se traduce en pendientes de crecimiento mayores en el índice de precios. Es interesante observar como las zonas intermedias son las que mejor se comportan en el resto de provincia, probablemente porque incluyen las metropolitanas residenciales de ingresos más altos, que como se verá posteriormente, crecen de forma importante.

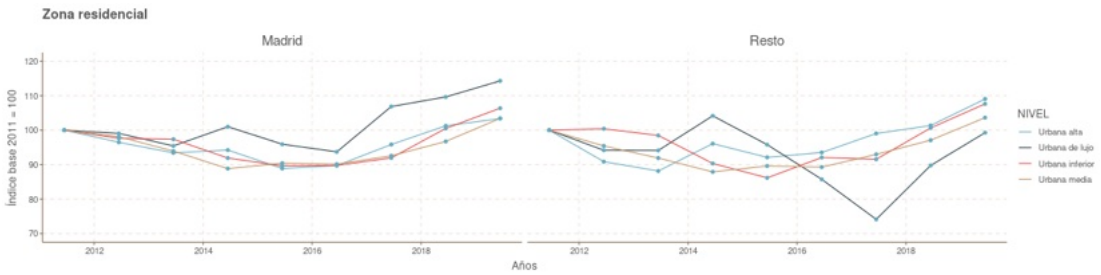
Figura 8.10. Índice de precio del alquiler desglosado por densidad de población



Fuente: elaboración propia.

La Figura 8.11 ilustra cómo las zonas de mayor nivel adquisitivo (lujo) presentan un comportamiento casi acíclico, con escasas caídas y un rápido crecimiento (particularmente para la ciudad de Madrid). Por otro lado, los demás estratos exhiben un comportamiento similar entre sí.

Figura 8.11. Índice de precio del alquiler desglosado por tipo de zonas residencial



Fuente: elaboración propia.

El desglose de los índices por características demográficas se presenta en la Tabla 8.9. Se confirma que las áreas densamente pobladas, los municipios más grandes y las zonas urbanas de lujo en Madrid experimentaron un crecimiento más pronunciado en los precios de las viviendas en comparación con otras áreas y categorías. Estos resultados parecen confirmar la hipótesis inicial de que las áreas con mayor población tienen un mejor comportamiento que los municipios

pequeños, al menos en términos de recuperación de precios de la vivienda.

Tabla 8.9. Índices de precios desglose funcional por tipo de zona

Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Densidad de población	Densamente poblada	Madrid	100	97.8	94.1	89.9	90.4	90.2	93.0	97.4	103.6
		Resto	100	94.7	92.0	89.3	90.0	90.7	94.4	99.6	106.5
	Diseminada	Resto	100	97.9	94.2	86.3	90.4	84.1	87.7	88.8	93.8
	Intermedia	Resto	100	99.9	94.6	90.7	94.7	95.3	100.3	103.9	112.9
Tamaño de municipio	10.000 o más y menos de 20.000 habitantes	Resto	100	98.0	89.5	84.3	84.1	82.0	84.1	87.8	92.4
	20.000 o más y menos de 50.000 habitantes	Resto	100	94.3	90.0	86.0	87.9	87.8	88.8	92.1	99.5
	50.000 o más y menos 100.000 habitantes	Resto	100	92.6	89.1	87.0	91.9	92.6	98.3	103.4	110.9
	Menos de 10.000 habitantes	Resto	100	97.2	93.8	85.4	87.7	82.1	85.0	86.4	91.6
	Municipio de 100.000 habitantes o más	Madrid	100	97.8	94.1	89.9	90.4	90.2	93.0	97.4	103.6
		Resto	100	96.6	94.0	90.9	87.6	87.9	90.1	94.8	100.6
Zona residencial	Urbana alta	Madrid	100	96.5	93.4	94.2	88.9	89.7	95.9	101.3	103.4
		Resto	100	90.9	88.2	96.1	92.1	93.5	99.0	101.4	109.0
	Urbana de lujo	Madrid	100	99.1	95.4	101.0	95.9	93.7	106.9	109.6	114.3
		Resto	100	94.2	94.1	104.2	95.8	85.7	74.1	89.8	99.3
	Urbana inferior	Madrid	100	97.6	97.4	91.9	89.7	89.8	92.0	100.5	106.4
		Resto	100	100.4	98.5	90.3	86.2	92.1	91.6	100.6	107.6
	Urbana media	Madrid	100	98.1	93.9	88.8	90.4	90.1	92.6	96.7	103.4
		Resto	100	95.4	91.9	87.9	89.6	89.3	93.0	97.1	103.6

Fuente: elaboración propia

Los datos indican que las áreas densamente pobladas muestran menores caídas y mayores crecimientos en los precios. Por ejemplo, estás áreas en Madrid experimentaron una recuperación en los precios de las viviendas, con un incremento del índice desde 89,9 en 2014 hasta 103,6 en 2019. De manera similar, el índice en el resto de la Comunidad de áreas similares también aumentó sustancialmente durante el mismo periodo, pasando de 89,3 en 2014 a 106,5 en 2019.

En cuanto al tamaño de los municipios, se observan diferentes tendencias a lo largo del periodo analizado en función del área de análisis. Aquellos con una población de entre 10.000 y 20.000 habitantes en el resto de la Comunidad registraron un aumento en el índice del 84,3 en 2013 a 92,4 en 2019. Por otro lado, los municipios de más de 100.000 habitantes en Madrid experimentaron un incremento sustancial en el índice, pasando de 89,9 en 2014 a 103,6 en 2019.

En el resto de municipios de este tamaño, el índice también aumentó, pero de manera más moderada, pasando de 90,9 en 2013 a 100,6 en 2019.

La evolución del índice de precios de vivienda también varió en función de la zona residencial. En las zonas urbanas de alta densidad en Madrid, el índice aumentó de 94,2 en 2014 a 103,6 en 2019, mientras que en el resto el crecimiento fue aún mayor, pasando de 96,1 en 2014 a 109 en 2019. En contraste, las áreas urbanas de menor densidad en Madrid registraron una evolución más moderada, con un aumento de 91,9 en 2014 a 106,4 en 2019, que en el resto el índice alcanzó niveles de 107,6 y 90,3 respectivamente.

Por otra parte, el comportamiento del índice en las zonas urbanas de lujo en Madrid mostró un crecimiento más acusado que el observado en otras áreas residenciales, pasando del 101 en 2014 al 114,3 en 2019. Por otro lado, para este estrato y periodo, el resto de la Comunidad registró una disminución del índice de 104,1 al 99,3. La recuperación de los precios en zonas urbanas de lujo en la capital está relacionada con una persistente demanda de viviendas de alta gama.

Tabla 8.10. Índices de precios desglose funcional por instalaciones de la vivienda

Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Aire acondicionado	No	Madrid	100	97.7	93.9	89.1	89.8	90.0	91.8	95.9	102.2
		Resto	100	95.8	92.0	88.7	89.5	89.7	93.5	97.1	103.7
	Sí	Madrid	100	97.9	94.3	91.0	91.2	90.6	94.4	99.1	105.1
		Resto	100	94.7	92.2	88.1	90.4	89.5	93.0	98.3	104.8
Parking	No	Madrid	100	98.0	93.7	89.0	89.2	89.1	91.7	96.2	102.3
		Resto	100	95.7	91.6	87.9	88.3	88.8	91.9	95.6	101.5
	Sí	Madrid	100	96.5	94.7	92.4	94.7	92.3	95.6	98.6	105.0
		Resto	100	95.6	93.5	90.7	93.6	92.1	97.2	102.5	110.4
Piscina	No	Madrid	100	97.9	93.9	89.3	89.6	89.8	92.5	96.9	103.2
		Resto	100	95.9	91.9	87.7	89.2	89.1	92.2	95.7	102.3
	Sí	Madrid	100	96.7	94.8	92.9	94.6	91.1	95.3	98.6	103.5
		Resto	100	94.7	93.0	91.4	92.4	91.8	97.2	102.9	109.8
Trastero	No	Madrid	100	97.8	93.7	89.3	89.7	89.7	92.4	96.8	103.0
		Resto	100	95.4	91.8	88.7	89.4	89.6	93.1	97.0	103.1
	Sí	Madrid	100	97.5	95.3	92.3	94.0	90.9	94.7	98.1	104.3
		Resto	100	95.8	93.4	89.0	92.0	90.7	94.9	100.2	108.4

Fuente: elaboración propia

El desglose de los índices en función de las instalaciones de la vivienda puede

aportar una perspectiva más ajustada a las tendencias y flujos del mercado inmobiliario en función de las preferencias del consumidor. La Tabla muestra 8.10 los valores del índice en función de si la vivienda dispone de 4 elementos: parking, la piscina, trastero y aire acondicionado. Se aprecia que existen variaciones significativas en función de la característica de estudio y el ámbito geográfico, que, en particular, se acentúan para resto de la Comunidad de Madrid. De los cuatros tipo de equipamiento el garaje y la piscina son las características que presentan una mayor divergencia de valores.

La presencia de aire acondicionado en la vivienda no muestra diferencias significativas en la evolución de los precios, dependiendo de si la vivienda cuenta o no con dicho equipamiento. Por ejemplo, las diferencias son menores a dos puntos, tanto en Madrid como en el resto de la Comunidad, en los años 2013 y 2019.

En cuanto a las viviendas con piscina en Madrid, esta característica resulta irrelevante en la evolución del precio. Se observa que el índice de precios en 2019 es de 103,5 para viviendas con piscina y de 103,2 para aquellas sin ella. Sin embargo, en el resto de la Comunidad sí existe una diferencia considerable: el índice de precios en 2019 en viviendas con piscina es de 109,7, en contraste con 102,3 en las viviendas sin piscina.

En relación al trastero, el índice de precios en 2019 para viviendas con trastero en Madrid era de 104,3, mientras que para aquellas sin trastero era de 103. En el resto de la Comunidad, se observa un incremento mayor en comparación con Madrid: el valor del índice en 2019 era de 108,4 para las viviendas con trastero y de 103,1 para las viviendas sin él.

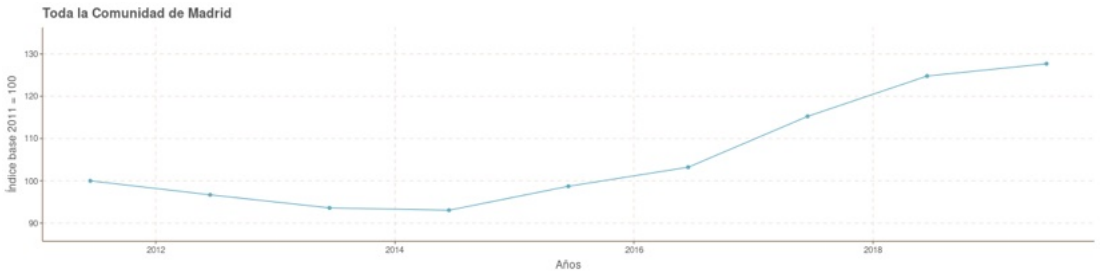
Finalmente, en lo que respecta al factor de estacionamiento, se observa un comportamiento similar al del trastero. Para las viviendas con estacionamiento en Madrid, el valor del índice en 2019 era de 104,9, mientras que para las viviendas sin él era de 102,3. En el resto de la Comunidad, el comportamiento es más extremo: el valor del índice en 2019 en viviendas con estacionamiento era de 110,4, mientras que en las viviendas sin estacionamiento era de 101,7.

8.3.3.2 Análisis de las series de oferta (desglose funcional)

En este epígrafe, se realiza un análisis análogo al del índice de precios de alquiler, pero esta vez desde el ángulo de la oferta. La principal diferencia que se aprecia en la Figura 8.12 es la caída de precios desde el año base hasta el año 2015, en lugar del 2016 como en el caso de los precios de mercado. Esto es esperable, ya que el precio de mercado es en cierta medidas un indicador retrasado del precio

de la oferta.

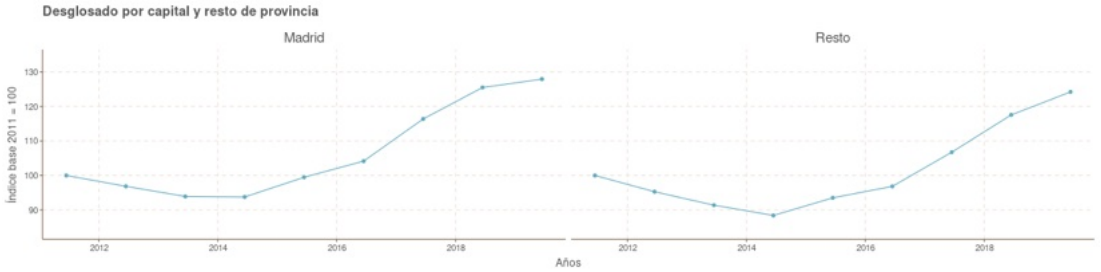
Figura 8.12. Índice de precio de oferta general



Fuente: elaboración propia.

Las diferencias por macrozonas, muestran como la capital comienza la recuperación de precios antes, y alcanza un periodo de saturación de precios alrededor de 2019 (Figura 8.13). También se aprecia como el resto de provincia tiene un retraso temporal con tasas de crecimiento más pronunciadas en el periodo final de la serie, a partir de 2016. Lo cual estaría relacionado con un mayor nivel de liquidez del mercado inmobiliario de la capital, comparado con el resto de áreas.

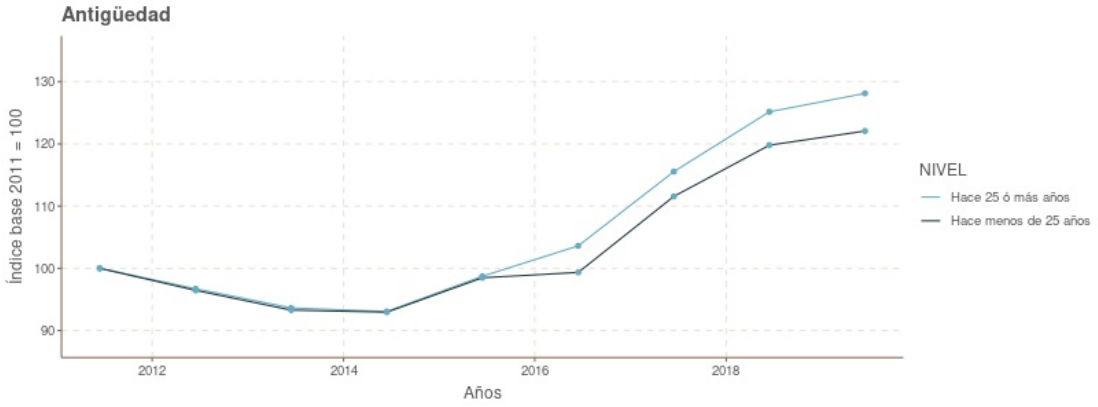
Figura 8.13. Índice de precio de oferta desglosado por Madrid o resto de provincia general



Fuente: elaboración propia.

Si atendemos a la descomposición funcional en detalle, el comportamiento es análogo al de mercado, los precios inmuebles más modernos tienen un mejor desempeño que en los más antiguos, tanto en la capital como en el resto de zonas, como se ve en la Figura 8.14. En general, se comprueba como las series desglosadas de oferta no muestran las discontinuidades puntuales de algunos índices de alquiler.

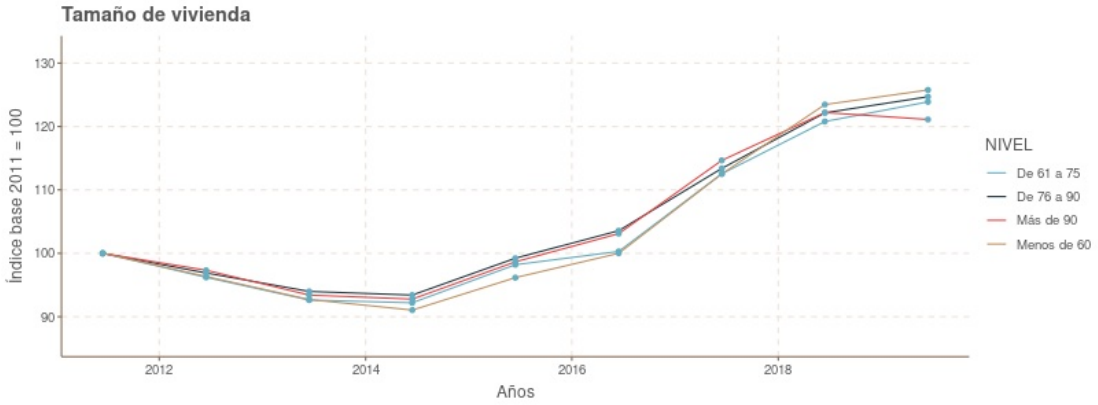
Figura 8.14. Índice de precio de oferta desglosado por año de construcción



Fuente: elaboración propia.

Los diversos rangos de superficie de vivienda no muestran diferencias apreciables, al contrario del caso de mercado, siendo todas las series muy similares en comportamiento, como se aprecia en la Figura 8.15. Se puede destacar como en este caso, las viviendas de superficie intermedia son las que más crecen en lugar de aquellas con mayor superficie, como se veía en el caso del precio de mercado.

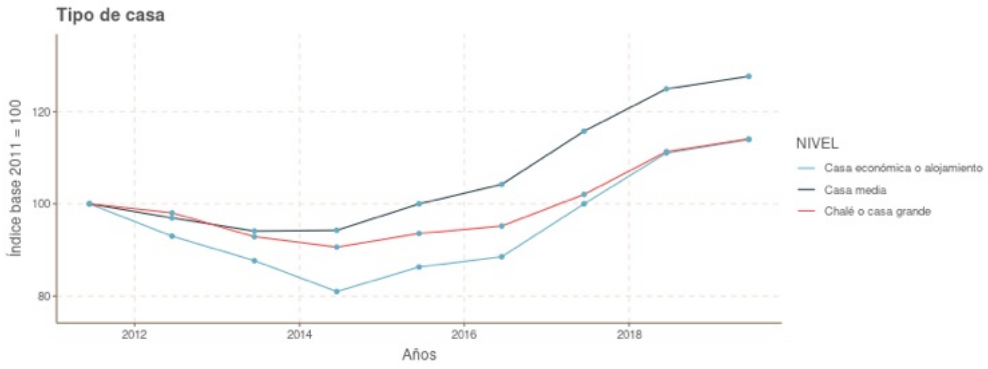
Figura 8.15. Índice de precio de oferta desglosado por superficie útil



Fuente: elaboración propia.

En cuanto a la tipología, las viviendas de tipo medio ofrecen el mejor comportamiento, al contrario del precio de mercado donde eran las unifamiliares, como se aprecia en Figura 8.16. Aunque en ambos casos, las viviendas más económicas son las que tienen una caída más pronunciada en torno al 2015. Para este último tipo de inmuebles, la serie de precios de las viviendas unifamiliares es mucho más estable, al mitigarse la inestabilidad de los índices ante muestras pequeñas y heterogéneas (Goh *et al.*, 2012).

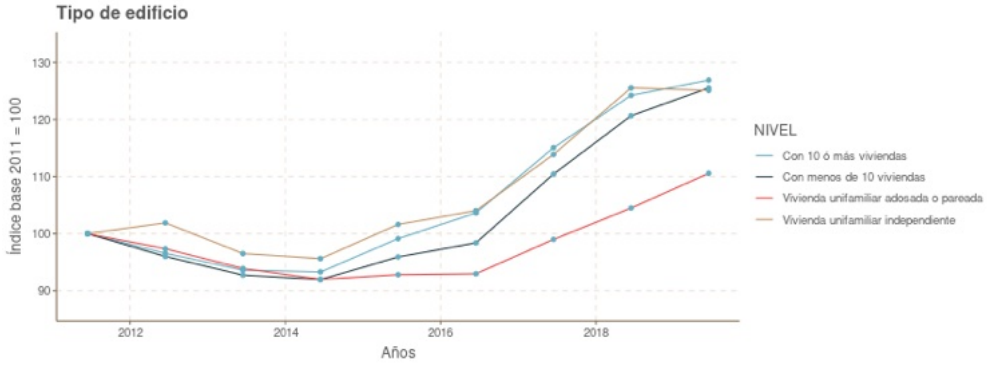
Figura 8.16. Índice de precio de oferta desglosado por tipo de vivienda



Fuente: elaboración propia.

Se observa que el desglose por tipo de edificio es consistente con los resultados por tipo de casa: mejor comportamiento de las viviendas intermedias y peor de las unifamiliares, como se ve en la Figura 8.17.

Figura 8.17. Índice de precio de oferta desglosado por tipo de edificio



Fuente: elaboración propia.

La evolución de los índices de precios de la vivienda basados en las características constructivas se analiza a través de la Tabla 8.11. Los datos revelan el fuerte aumento en todos los índices de precios desde 2013 hasta 2019, tanto en Madrid como en el resto de la Comunidad de Madrid, con diferencias en función del estrato y el área. Por ejemplo, se aprecia un incremento del precio en vivienda nueva fuera de Madrid o una mayor apreciación, en general, en las viviendas con menor superficie.

Tabla 8.11. Índices de precios desglose funcional por estructura

Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Antigüedad	Hace 25 ó más años	Madrid	100	96.9	93.9	93.7	99.5	104.5	116.6	125.8	128.4
		Resto	100	94.5	90.7	87.4	91.5	95.4	104.0	114.7	121.9
	Hace menos de 25 años	Madrid	100	96.0	93.2	93.8	98.2	98.9	110.8	117.2	119.0
		Resto	100	97.3	93.3	91.7	98.9	99.9	112.3	122.8	127.4
Número de habitaciones	1 o 2 habitaciones	Madrid	100	96.7	93.0	91.9	97.5	101.5	113.1	123.0	123.8
		Resto	100	94.9	91.5	90.1	94.2	96.9	108.1	118.5	123.7
	3 habitaciones	Madrid	100	95.6	91.1	90.3	96.7	97.1	109.2	118.5	120.6
		Resto	100	95.3	90.2	85.9	89.9	95.0	105.1	116.6	124.8
	4 habitaciones	Madrid	100	97.7	92.8	91.1	98.3	98.7	113.8	121.7	122.6
		Resto	100	99.8	93.5	89.7	94.8	98.0	109.7	122.0	128.1
	5 o más habitaciones	Madrid	100	96.6	91.8	91.9	93.3	102.9	109.2	116.0	114.5
		Resto	100	95.9	87.9	83.7	92.4	95.6	100.1	104.0	113.2
Tamaño de vivienda	De 61 a 75	Madrid	100	96.4	93.3	93.5	100.2	101.1	112.8	119.9	122.5
		Resto	100	95.7	91.5	90.0	94.8	98.8	112.1	123.1	128.6
	De 76 a 90	Madrid	100	97.3	95.3	95.9	103.4	108.4	118.4	124.4	123.9
		Resto	100	96.1	92.0	90.5	94.9	98.3	107.2	118.8	127.0
	Menos de 60	Madrid	100	96.4	92.8	91.1	96.3	100.2	112.7	123.6	125.8
		Resto	100	94.7	91.3	90.5	92.0	96.6	107.5	118.5	126.9
	Más de 90	Madrid	100	96.7	93.3	94.0	100.6	105.1	115.7	121.6	118.4
		Resto	100	98.9	93.6	90.8	96.1	99.7	113.2	123.1	127.0
Tipo de casa	Casa económica o alojamiento	Madrid	100	93.5	86.9	79.5	84.0	86.3	100.3	110.8	115.9
		Resto	100	91.2	89.4	83.5	88.5	90.7	99.7	111.3	112.4
	Casa media	Madrid	100	97.0	94.4	94.7	100.5	105.1	116.7	125.5	127.9
		Resto	100	96.0	91.8	90.9	96.0	98.1	108.4	118.7	124.9
	Chalé o casa grande	Madrid	100	93.3	92.0	89.0	83.6	91.2	90.7	101.4	109.7
		Resto	100	99.9	93.1	91.0	94.8	95.8	104.1	112.0	114.3
Tipo de edificio	Con 10 ó más viviendas	Madrid	100	96.6	93.9	93.8	99.7	104.3	115.9	124.6	127.0
		Resto	100	95.6	91.0	89.0	95.0	98.1	107.7	119.2	124.6
	Con menos de 10 viviendas	Madrid	100	97.5	92.1	92.5	94.3	94.8	107.2	119.2	121.6
		Resto	100	95.1	93.1	91.7	96.6	99.6	111.5	121.1	127.5
	Vivienda unifamiliar adosada o pareada	Madrid	100	93.3	92.0	89.0	83.6	91.2	90.7	101.4	109.7
		Resto	100	99.1	94.4	92.7	94.1	93.3	100.8	104.7	110.6
	Vivienda unifamiliar independiente	Resto	100	101.9	96.5	95.6	101.6	104.0	113.9	125.5	125.1

Fuente: elaboración propia

En base a la antigüedad de las casas, se puede observar una tendencia decreciente hasta 2013, tanto en Madrid como en el resto de la Comunidad. Sin embargo, en 2019, los índices mostraron un aumento en ambas áreas, con incrementos más fuertes en la capital que el resto de provincia, llegando a 128,4 en Madrid para las casas con 25 años o más, y 127,4 en el resto de la Comunidad para casas con menos de 25 años. Este comportamiento estaría relacionado con una demanda de inmuebles nuevos no satisfecha, por la escasa actividad de construcción de este tipo de viviendas en los años 2008-2020.

En relación con el número de habitaciones, se observa cierta variabilidad en los índices entre 2013 y 2019. En 2013, las casas con 1 o 2 habitaciones presentaron índices similares entre Madrid (93,0) y el resto de la Comunidad (91,5). Aunque en 2019 ambos índices aumentaron, la diferencia entre ellos se mantuvo casi constante, siendo 123,8 en Madrid y 123,7 en el resto de la Comunidad. Sin embargo, en la categoría de viviendas con 5 o más habitaciones, el índice en Madrid disminuyó de 91,8 en 2013 a 114,5 en 2019, lo que contrasta con el aumento en el índice en el resto de la Comunidad (87,9 en 2013 a 113,2 en 2019). El mayor crecimiento en el segmento de pisos pequeños está motivado los cambios demográficos sucedidos en dicho periodo, y que se resumen en el aumento de hogares compuestos por una o dos personas y la caída de hogares de 4 o más miembros.

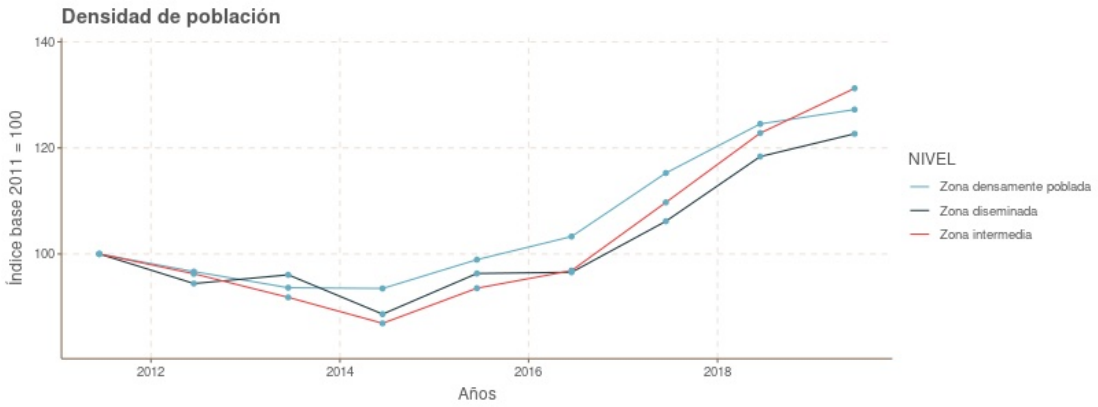
Para el criterio de tamaño de las viviendas, encontramos una tendencia creciente en los índices. Las viviendas de 76 a 90 metros cuadrados, presentaron un aumento en los índices en ambas áreas, siendo de 95,3 (Madrid) y 92,0 (Resto) en 2013, y 123,9 (Madrid) y 127 (Resto) en 2019. Para las viviendas de menos de 60 metros cuadrados, se observó un incremento en los índices de 92,8 en 2013 a 125,8 en 2019 en Madrid y de 91,3 en 2013 a 126,9 en 2019 en el resto de la Comunidad. De nuevo, los cambios demográficos motivan una mayor demanda en las viviendas con menor superficie.

Finalmente, en cuanto al tipo de vivienda, se encuentra una tendencia alcista en los índices de precios de 2013 a 2019. La casa media en Madrid mostró un aumento en su índice de 94,4 en 2013 a 127,0 en 2019, mientras que en el resto de la Comunidad, el índice se incrementó de 91,8 a 124,9 en el mismo período. Asimismo, la vivienda unifamiliar adosada o pareada en Madrid presentó un incremento en su índice de 92,0 en 2013 a 109,7 en 2019, mientras que en el resto de la Comunidad, el índice aumentó de 94,4 en 2013 a 110,6 en 2019.

Las diferencias de los índices de precios, según la densidad de población de la zona, son mínimas, aunque similares a lo visto para los precios de mercado, con un mejor desempeño de las zonas de tipo intermedio comparadas con las zonas

diseminadas, como muestra la Figura 8.18.

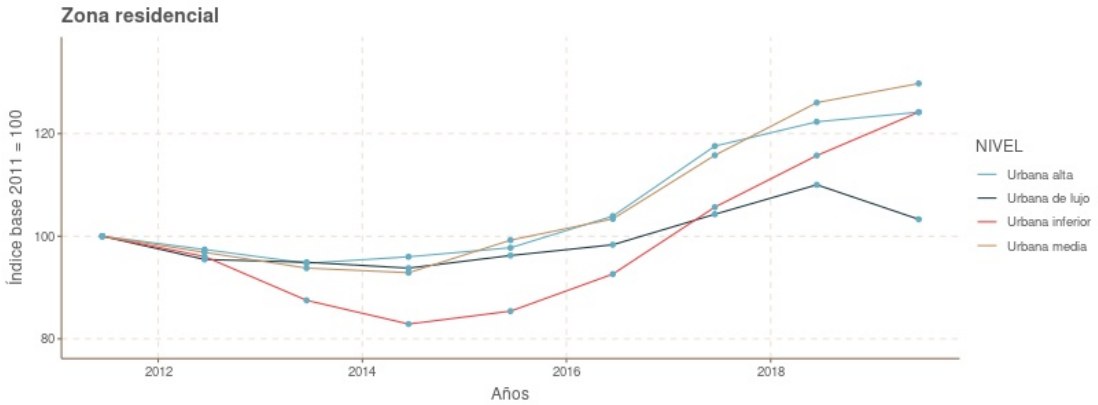
Figura 8.18. Índice de precio de oferta desglosado por densidad de población



Fuente: elaboración propia.

En el desglose por ingresos, los precios de oferta tienen un comportamiento sensiblemente diferente a los de mercado. Se aprecia un comportamiento más extremo en las zonas de menor ingresos (urbana inferior), con unos descensos más pronunciados y una recuperación más brusca. Las zonas de lujo muestran cierta saturación y decrecimiento en Madrid a partir de 2018, como indica la Figura 8.19.

Figura 8.19. Índice de precio de oferta desglosado por tipo de zona residencial



Fuente: elaboración propia.

A continuación, se presenta una comparación de los índices de precios en 2013 y 2019, categorizados por densidad de población, tamaño de municipio y zona residencial, en la Tabla 8.12. En general, las series muestran una tendencia creciente desde 2014, cuyos aumentos más significativos suceden en áreas con densidades de población intermedias, municipios con 10.000 a 20.000 habitantes y zonas urbanas inferiores. Por otra parte, las zonas urbanas de lujo en Madrid

muestran un descenso moderado en el índice de precios durante el período analizado.

Tabla 8.12. Índices de precios desglose funcional por tipo de zona

Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Densidad de población	Densamente poblada	Madrid	100	96.7	93.9	93.8	99.3	103.8	115.9	124.9	127.2
		Resto	100	96.1	92.1	91.2	96.4	99.8	110.4	121.0	127.0
	Diseminada	Resto	100	94.4	96.1	88.7	96.3	96.5	106.2	118.4	122.7
	Intermedia	Resto	100	96.3	91.8	86.9	93.5	96.9	109.7	122.8	131.2
Tamaño de municipio	10.000 o más y menos de 20.000 habitantes	Resto	100	96.9	90.1	83.4	90.4	91.6	97.4	108.7	115.8
	20.000 o más y menos de 50.000 habitantes	Resto	100	96.0	90.9	88.2	92.5	95.3	101.9	115.8	121.7
	50.000 o más y menos 100.000 habitantes	Resto	100	95.2	91.4	91.0	97.5	99.2	113.4	123.1	126.5
	Menos de 10.000 habitantes	Resto	100	93.0	93.9	86.6	89.6	90.6	98.9	109.8	113.9
	Municipio de 100.000 habitantes o más	Madrid	100	96.7	93.9	93.8	99.3	103.8	115.9	124.9	127.2
		Resto	100	95.2	89.9	86.2	89.8	94.1	101.3	112.0	121.8
Zona residencial	Urbana alta	Madrid	100	97.4	94.5	95.7	97.6	104.0	117.8	122.7	124.2
		Resto	100	97.9	96.9	98.9	99.5	103.4	115.8	119.8	123.8
	Urbana de lujo	Madrid	100	95.6	92.3	91.3	86.5	93.6	100.7	105.0	101.1
		Resto	100	95.4	95.7	94.5	98.5	99.8	105.7	111.1	116.2
	Urbana inferior	Madrid	100	96.5	87.7	83.0	85.2	90.9	108.3	115.8	121.0
		Resto	100	94.7	86.9	82.7	85.8	95.3	100.9	115.5	139.9
	Urbana media	Madrid	100	97.0	94.1	93.5	100.1	104.4	117.0	126.9	130.2
		Resto	100	95.3	91.7	88.7	93.7	96.4	106.0	117.2	122.7

Fuente: elaboración propia

En términos de densidad de población, las zonas densamente pobladas muestran un considerable aumento en el índice de precios, para el caso de la capital, pasa del 93,9 en 2013 al 127,2 en 2019. Por otra parte, la densidad de población en zonas intermedias experimentan un aumento considerable del índice, pasando de 91,8 en 2013 a 131,2 en 2019, convirtiéndose en una de las áreas que más crecimiento experimentan en ese período.

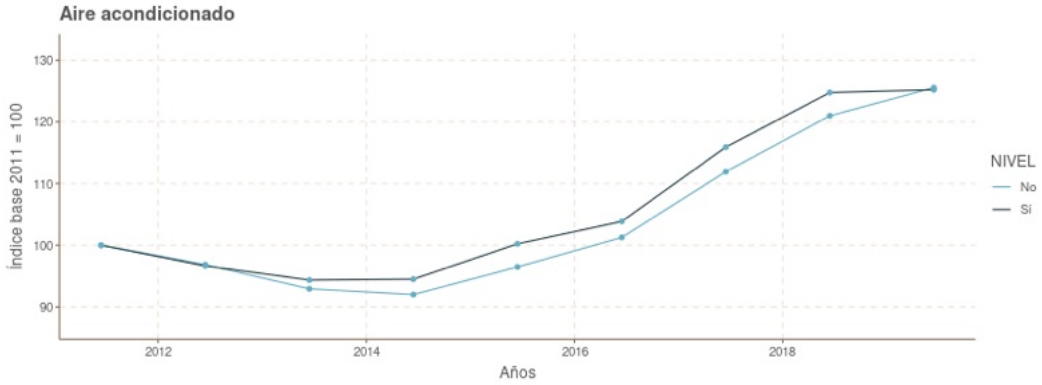
En cuanto al tamaño del municipio, la tendencia general es un aumento del índice de precios de oferta en todos los rangos analizados. Destacan los municipios con 10.000 o más y menos de 20.000 habitantes que experimentan un aumento del índice de 90,1 en 2013 a 115,8 en 2019, y los municipios de 100.000 habitantes o más del resto de Comunidad de Madrid, donde el índice crece de 89,9 en 2013 a 121,8 en 2019. En comparación, los municipios con menos de 10.000 habitantes

mantienen un crecimiento más modesto, con un índice que pasa de 93,9 en 2013 a 113,9 en 2019.

Finalmente, al considerar las zonas residenciales, se observa un considerable incremento en las urbanas inferiores en el resto de Comunidad, con el índice de precios de oferta aumentando del 86,9 en 2013 al 139,9 en 2019. En contraste, las áreas urbanas de lujo en Madrid experimentaron un descenso en el índice, pasando de 92,3 en 2013 a 101,1 en 2019. Las zonas residenciales urbanas medias en Madrid, por otro lado, mantienen un crecimiento equilibrado, con un índice que pasa de 94,1 en 2013 a 130,2 en 2019.

La evolución de precios en función los equipamientos en oferta y mercado tienen un funcionamiento similar, como se aprecia en la Figura 8.20.

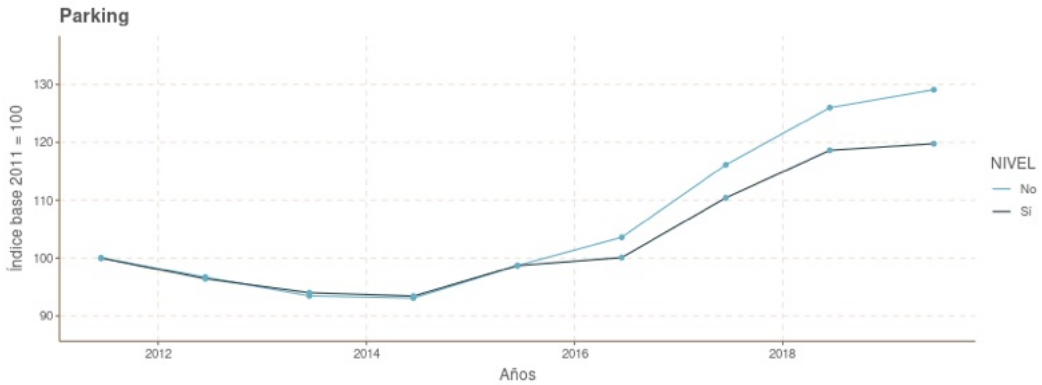
Figura 8.20. Índice de oferta por disponibilidad de aire acondicionado



Fuente: elaboración propia.

Sin embargo, existe una mayor diferencia en los crecimientos de las viviendas con garaje con respecto a las que no lo tienen en el caso de la capital, como se ve en la Figura 8.21.

Figura 8.21. Índice de oferta por disponibilidad de garaje



Fuente: elaboración propia.

El análisis del índice de oferta en función del equipamiento de la vivienda, mostrado en la Tabla 8.13, indica que ofrece más diferencias ha sido el aparcamiento, seguida de la presencia de piscina y trastero. Al igual que en el índice del alquiler, el aire acondicionado es un factor que revele diferencias en los precios. Lo cual se relaciona con que los compradores otorgan una mayor importancia a otros factores en cuanto a la valoración de las viviendas, o que el aire acondicionado ha pasado a ser un elemento casi habitual en muchas de ellas.

Tabla 8.13. Índices de precios desglose funcional por instalaciones de la vivienda

Factor	Nivel	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Aire acondicionado	No	Madrid	100	97.0	93.2	92.6	97.0	102.2	112.7	121.6	126.0
		Resto	100	95.9	91.4	88.7	93.2	96.2	106.6	116.5	121.4
	Sí	Madrid	100	96.8	94.6	94.9	100.6	104.4	116.7	125.0	125.2
		Resto	100	95.0	92.3	90.4	96.5	98.3	108.7	120.6	125.9
Parking	No	Madrid	100	96.9	93.8	93.9	99.5	104.5	117.2	126.8	129.4
		Resto	100	95.0	90.3	86.9	91.8	95.3	104.8	114.7	122.0
	Sí	Madrid	100	96.3	93.8	93.6	98.6	99.8	108.6	113.8	114.4
		Resto	100	96.7	94.3	93.2	98.9	100.4	112.1	123.9	127.7
Piscina	No	Madrid	100	96.7	93.7	93.5	99.1	104.2	116.7	126.0	128.8
		Resto	100	94.6	90.2	86.6	91.9	94.8	104.6	114.9	122.7
	Sí	Madrid	100	97.2	95.1	95.8	100.0	101.6	110.3	116.2	115.2
		Resto	100	98.0	95.6	94.8	99.1	101.2	111.7	122.5	125.2
Trastero	No	Madrid	100	96.6	93.7	93.7	99.1	104.1	116.3	125.6	128.3
		Resto	100	95.1	90.6	88.2	93.4	96.3	107.1	117.4	123.5
	Sí	Madrid	100	97.1	94.5	94.3	100.1	101.1	111.4	117.0	117.7
		Resto	100	97.4	95.2	92.7	97.7	100.2	109.7	122.1	126.9

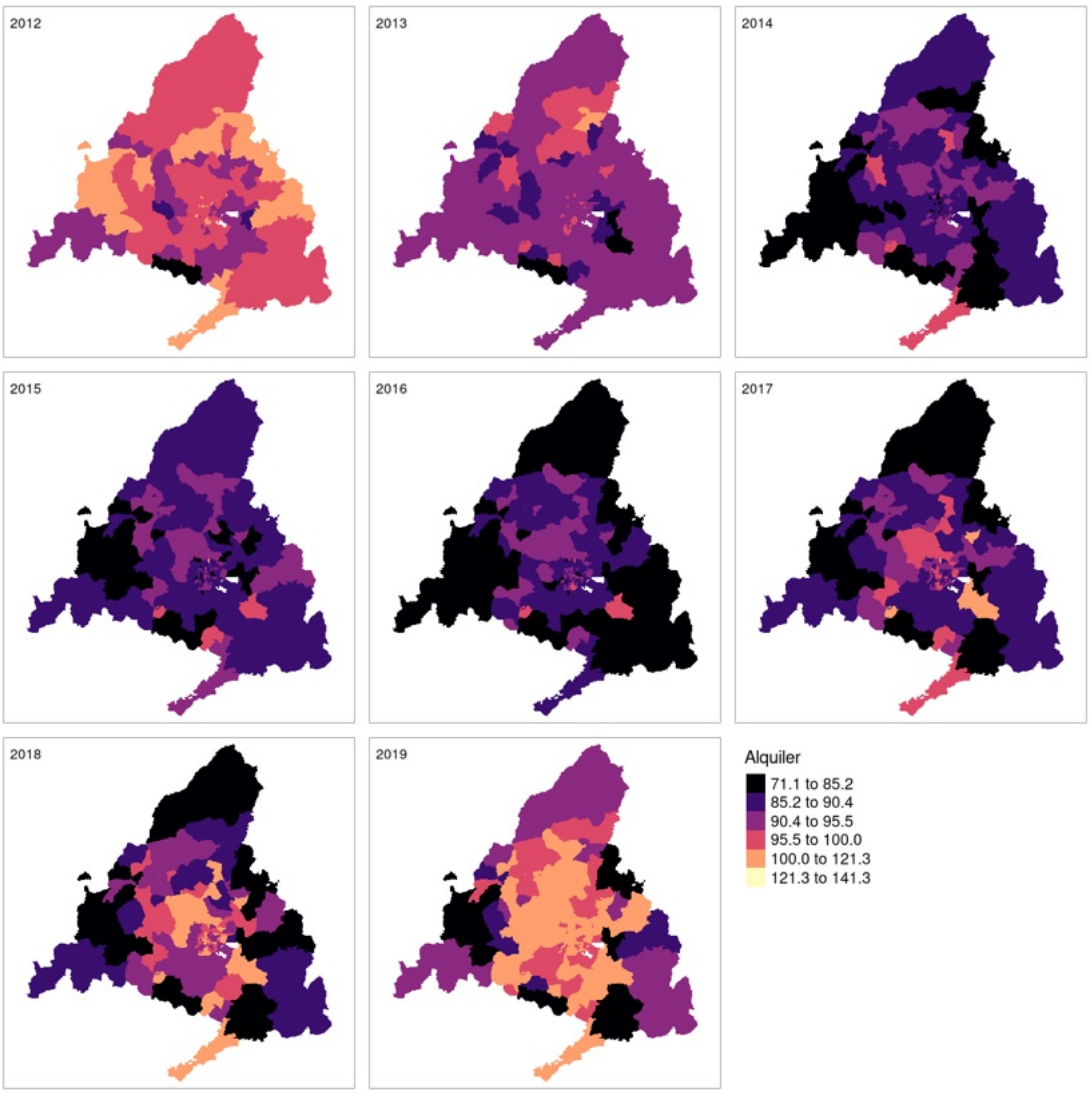
Fuente: elaboración propia

En relación al aparcamiento, los índices para viviendas sin garaje mostraban valores de 93,8 y 90,3 en Madrid y el resto de la región respectivamente, y de 93,8 y 94,3 cuando disponían de él. En cuanto a la piscina, se desprendía una diferencia más marcada entre viviendas con piscina (95,1 en Madrid y 95,6 en el resto) y sin ella (93,7 y 90,2). La situación para el trastero resulta equivalente, siendo de 93,7 y 90,6 en viviendas sin trastero y de 94,5 y 95,2 en viviendas con trastero, en la ciudad de Madrid y el resto de la provincia, respectivamente.

8.3.3.3 Análisis geográfico de las series

La evolución de los precios se muestra desigual a lo largo de la geografía de la Comunidad de Madrid, observándose de forma general como la zona central y más poblada es la primera en mostrar caídas de precios (Figura 8.22), y también es la primera zonas que muestra síntomas de recuperación. Por contra, las zonas periféricas y de tipo rural muestran una recuperación más lenta, presentando valores en 2019 inferiores a los del año base (2011).

Figura 8.22. Evolución del precio de mercado, todas las zonas

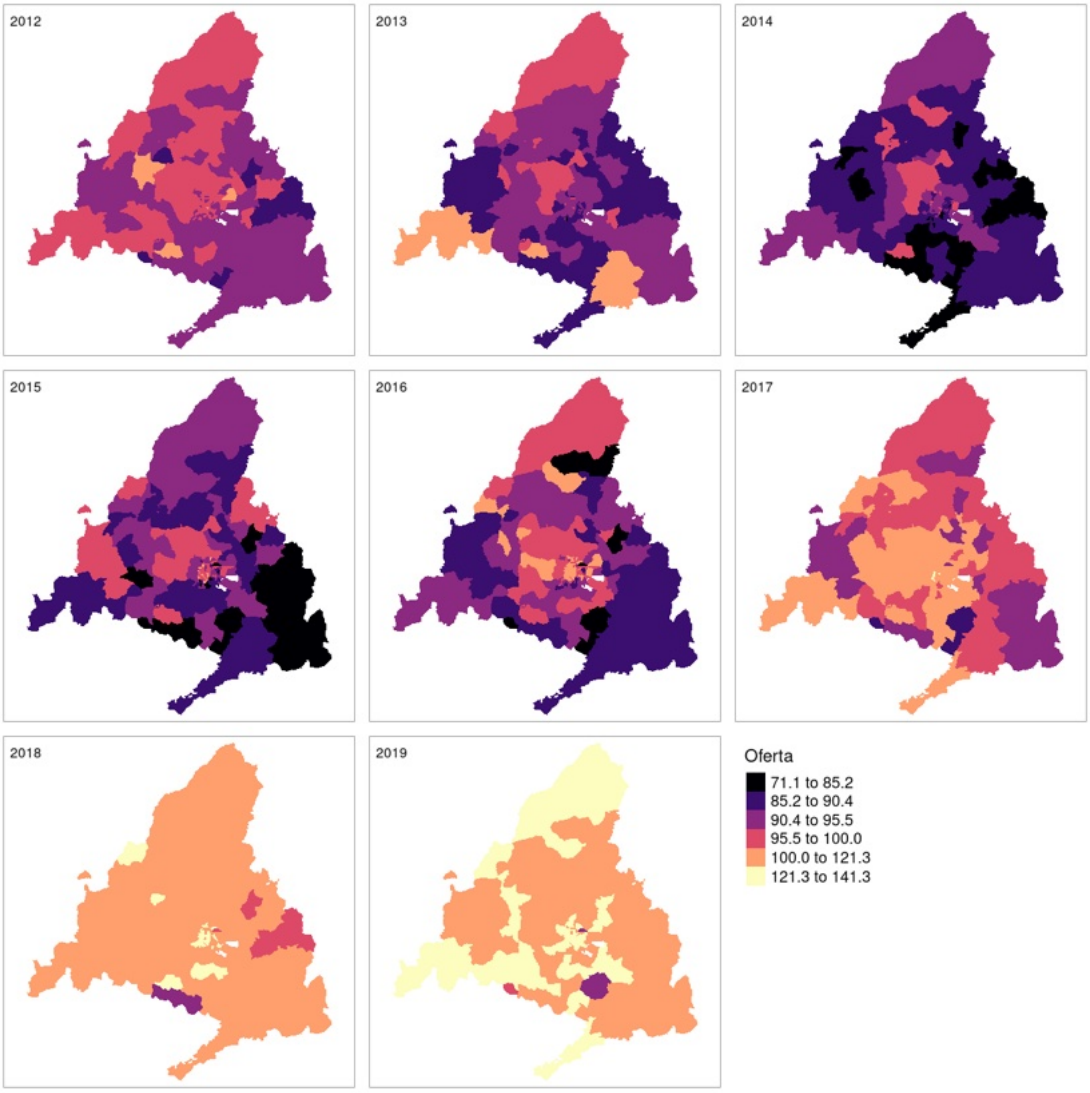


Fuente: elaboración propia.

De forma análoga, los precios de oferta muestran comportamiento similares a los de alquiler en el tiempo, como se ve en la Figura 8.23. Existen discrepancias, tales como que las zonas norte y oeste (sierra) experimentan caídas en oferta de menor intensidad que las observadas en los precios del alquiler. También se

observa dichas zonas, como junto con los barrios de Madrid, son las primeras en recuperarse. Esto es debido a que al ser en gran medida zonas de segundas residencias no estarían sujetas a las mismas dinámicas que las zonas urbanas (menor interés de los propietarios en alquilarlos a menor precio). Por otra parte, se muestra también que la oferta actúa como un indicador adelantado del precio del mercado (aunque con escalas de valores ligeramente diferentes).

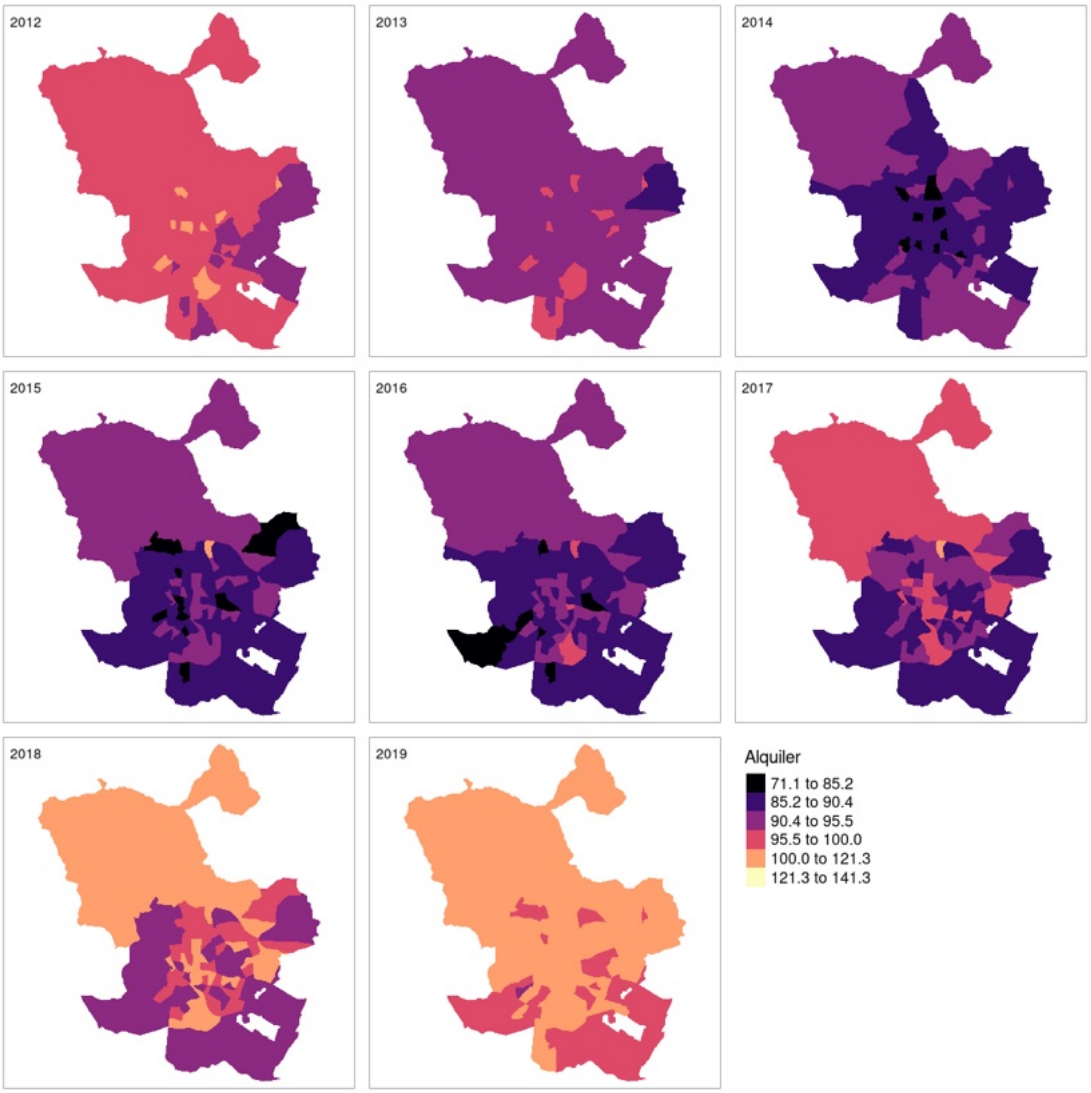
Figura 8.23. Evolución del precio de oferta, todas las zonas



Fuente: elaboración propia.

La capital es la primera zona en recuperarse y de las últimas en depreciarse, al ser zonas de muy alta demanda (independientemente del ciclo económico). Por contra, son aquellas zonas con menor poder adquisitivo las que tardan más tiempo en recuperarse.

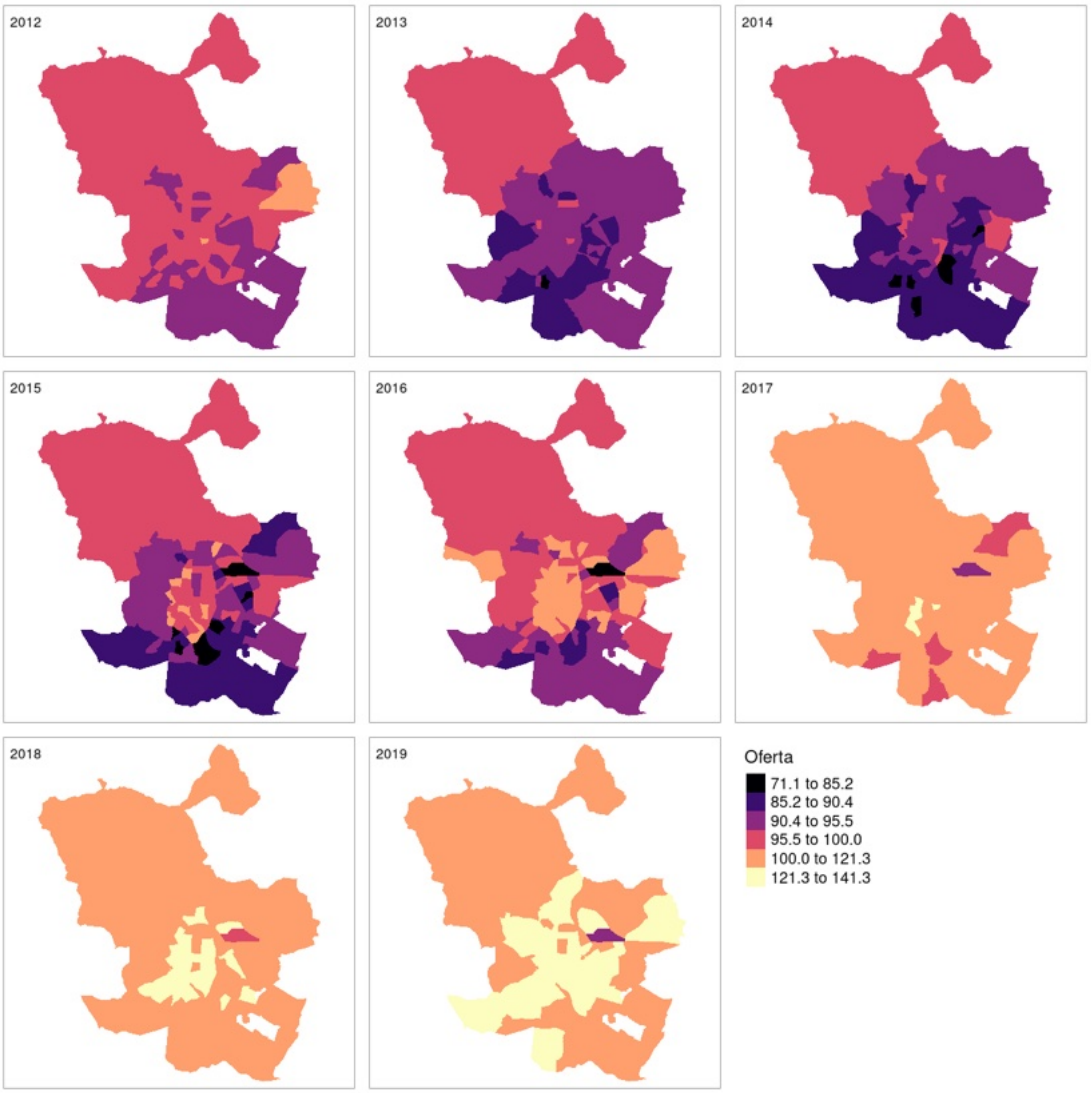
Figura 8.24. Evolución del precio de mercado, Madrid



Fuente: elaboración propia.

Como se apreciaba en toda la Comunidad, el precio de oferta en la ciudad (Figura 8.22), actúa como indicador adelantado del precio del alquiler (Figura 8.23), principalmente en el proceso de recuperación de precios, siendo la zona sur y este las zonas con una recuperación más lenta. Al igual que en el resto de la provincia, las zonas de mayor demanda (centro de la ciudad y eje del Paseo de la Castellana) son las que lideran la senda de recuperación de los precios.

Figura 8.25. Evolución del precio de oferta, Madrid



Fuente: elaboración propia.

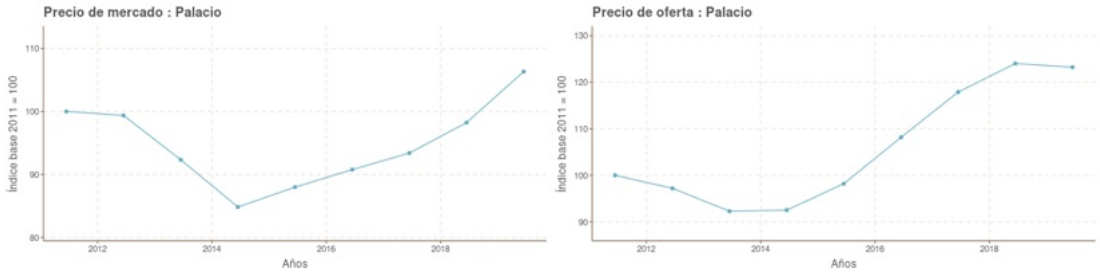
Para desarrollar un análisis geográfico en mayor en profundidad, se eligen una serie de municipios que representen distintas categorías de zona. Se segmenta por tipo de área y nivel de ingresos[10], que para el caso de Madrid capital son tres barrios[11] :

- Zona centro y turística: barrio Palacio.
- Zona residencial de alto poder adquisitivo, barrio de Vallehermoso.
- Zona residencial de medio-bajo poder adquisitivo, barrio de Orcasur-San Fermín.

[10]Se toman como referencia los indicadores de renta a barrios del municipio de Madrid, véase (Ayuntamiento de Madrid, 2022).
[11]Se han tomado los barrios de Palacio, Vallehermoso y Orcasur-San Fermín por ser zonas que representan diferentes submercados y cuentan con una alta representación, en número de inmuebles en la muestra.

Para el caso del barrio de Palacio (Figura 8.26), se observa que el índice generado es sensible a la composición muestral, y en el caso de 2014 muestra un cambio de tendencia abrupto. Sin embargo, también se aprecia que la recuperación comienza entre el año 2014 y 2015, iniciada un año antes desde la oferta. Además, la caída de precios es moderadamente gradual hasta 2014, cuando se invierte la tendencia hasta un nuevo punto de inflexión en el año 2018.

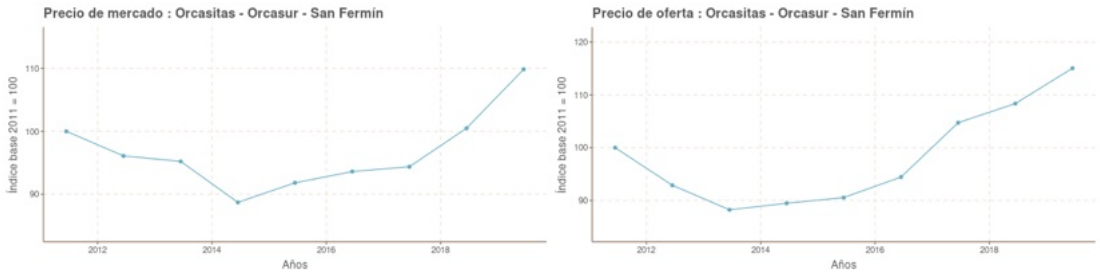
Figura 8.26. Índice en zona de turística (Madrid): Palacio



Fuente: elaboración propia.

En el caso del barrio de Orcasitas (Figura 8.27), los precios tienen una respuesta más lenta que el caso del centro, principalmente, por tener un nivel de demanda bajo, que se traduce en un crecimiento de precios en oferta más lento.

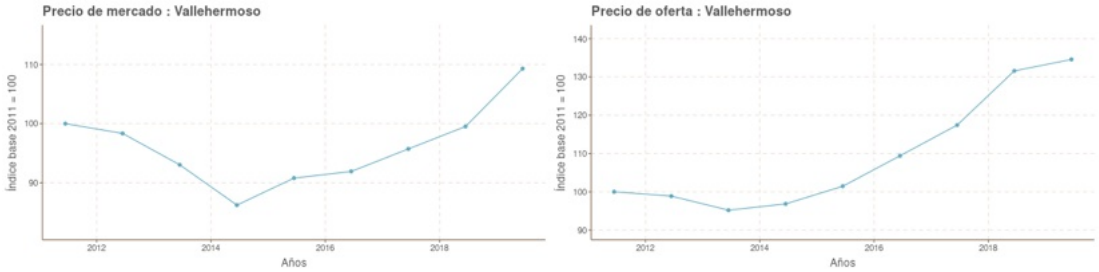
Figura 8.27. Índice en zona de ingresos medios-bajos (Madrid): Orcasur-San Fermín



Fuente: elaboración propia.

En el caso del barrio de Vallehermoso, no muestra ninguna caída abrupta de precios y la recuperación es similar a la que se ven en Palacio, véase la Figura 8.28. Se aprecia, también, el fenómeno de estancamiento de precios a partir de 2018.

Figura 8.28. Índice en zona de ingresos altos (Madrid): Vallehermoso



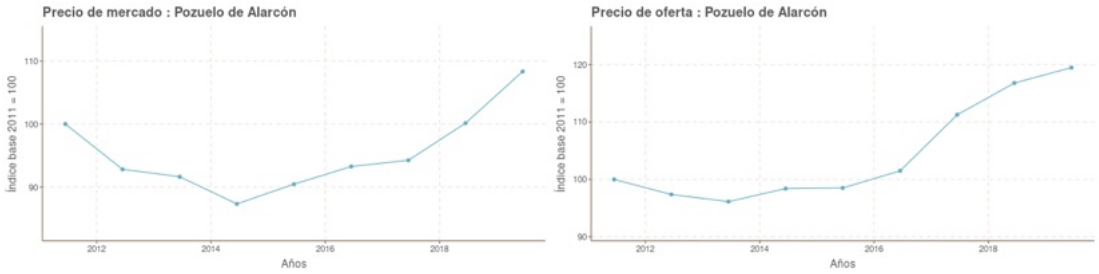
Fuente: elaboración propia.

En el área metropolitanta y zonas periféricas de la Comunidad de Madrid, se toman las siguientes tres zonas[12]:

- Municipio en zona metropolitana de altos ingresos: Pozuelo de Alarcón.
- Municipio en zona metropolitana de medios-bajos ingresos: Fuenlabrada.
- Municipio en zona rural: Área Sierra Norte.

En general, todas los municipios de la Comunidad ofrecen comportamientos en alquiler más moderados que en la ciudad. En el caso de Pozuelo, tras una caída de precios inicial, muestra un comportamiento plano hasta el año 2016, que se invierte en 2018 (Figura 8.29).

Figura 8.29. Índice en municipio de ingresos altos: Pozuelo de Alarcón

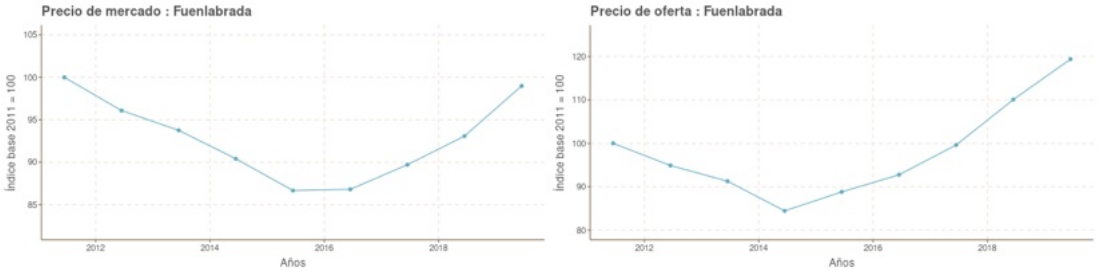


Fuente: elaboración propia.

Fuenlabrada, en cambio, como se puede ver en la Figura 8.30, sigue una caída progresiva hasta el año 2015, con una recuperación a partir del año 2016. En esta localidad, la transferencia de los precios de oferta al alquiler es más lenta que en Pozuelo, porque sus precios comienzan a crecer más tarde que en dicha localidad.

[12]Se han tomado los municipios de Pozuelo de Alarcón, Fuenlabrada y Área Sierra Norte representativas de los tres segmentos de mercado de estudio, y cuentan un número suficienteme de muestra.

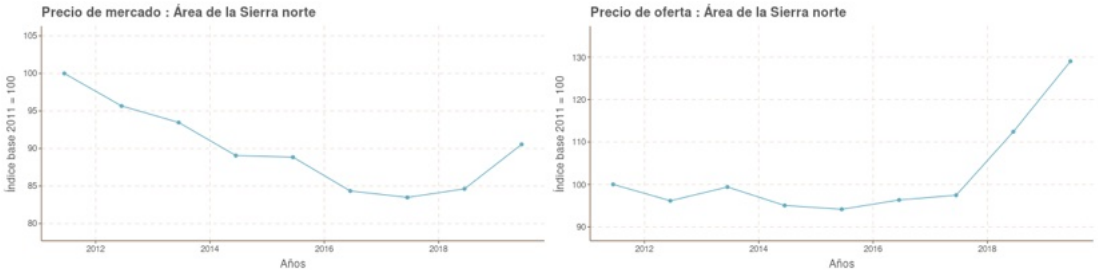
Figura 8.30. Índices generales por municipio, ingresos medios-bajos: Fuelabrada



Fuente: elaboración propia.

Por otra parte, la Zona Norte (Figura 8.31) no muestra un comportamiento tan extremo como las zonas metropolitanas. Se observa un retraso en la entrada de la fases crecientes de los precios de oferta y de alquiler que en los dos casos anteriores, motivado por el menor dinamismo inmobiliario de la zona.

Figura 8.31. Índices en municipio de tipo rural: Área Sierra Norte

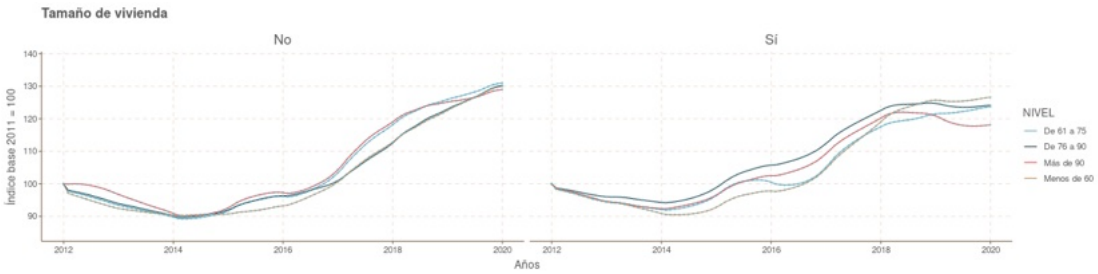


Fuente: elaboración propia.

8.3.3.4 Análisis de las series mensuales

A continuación se muestran varias de las series con frecuencia mensual. Los desgloses por rangos de superficie muestran las mismas características de tendencia que las series originales, como vemos en la Figura 8.32, para el caso de la evolución del precio de oferta.

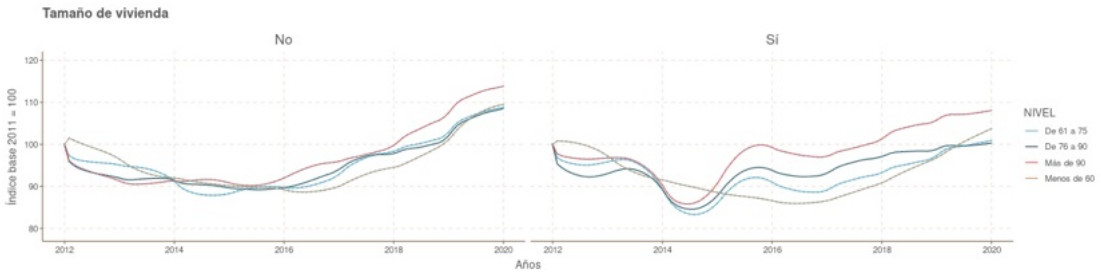
Figura 8.32. Índice de precio mensual de oferta desglosado por superficie útiles



Fuente: elaboración propia.

Los métodos de desagregación seleccionados tienen dificultad a la hora de generar series cuando el cambio interanual es importante, situación que se acentúa en el caso de las series de alquiler. Si las diferencias entre los años contiguos son menores, la aparición de discontinuidades en las curvas en menor, como se observa en la Figura 8.33.

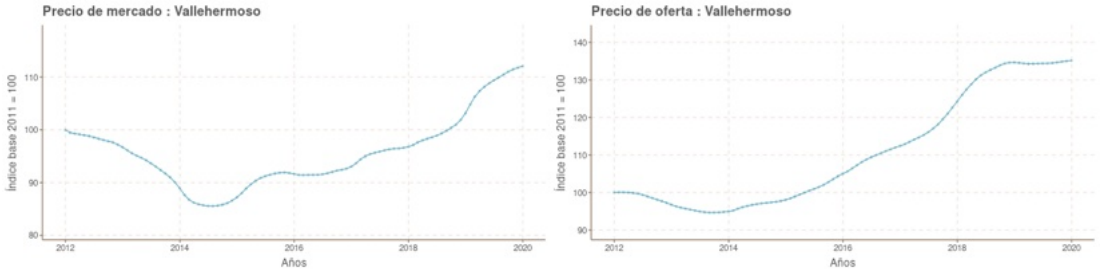
Figura 8.33. Índice de precio mensual de alquiler desglosado por superficie útiles



Fuente: elaboración propia.

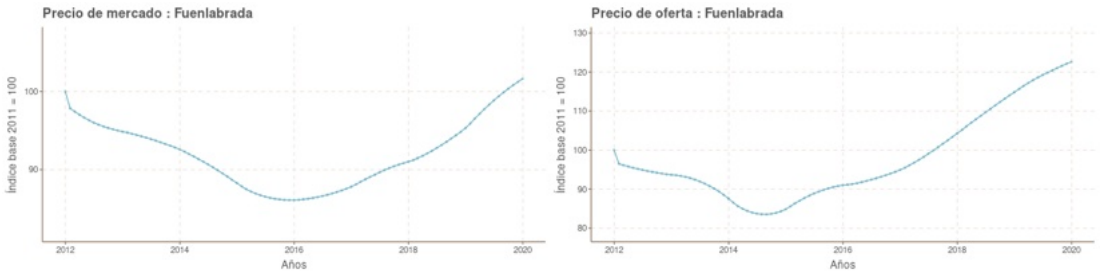
A continuación se muestran los desgloses geográficos analizados con frecuencia mensual, en las Figuras 8.34, 8.35 y 8.36. Se observa que el método de desagregación utilizado influye en la suavidad de las series generadas.

Figura 8.34. Índice mensual de alquiler, zona ingresos altos en Madrid: Vallehermoso



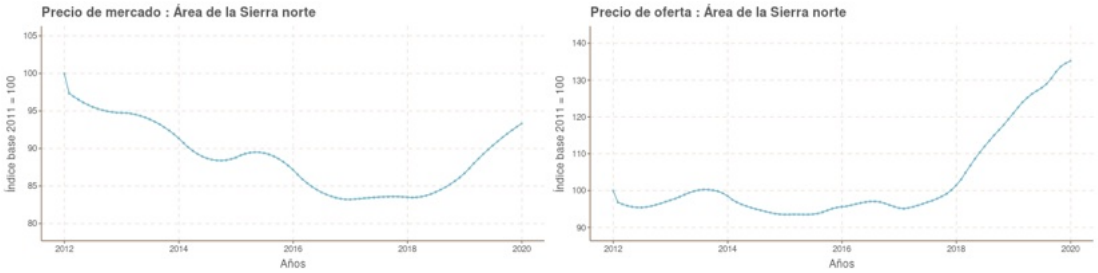
Fuente: elaboración propia.

Figura 8.35. Índice de alquiler mensual por municipio, ingresos medios-bajos: Fuenlabrada



Fuente: elaboración propia.

Figura 8.36. Índice de alquiler mensual en municipio rural: Área Sierra Norte



Fuente: elaboración propia.

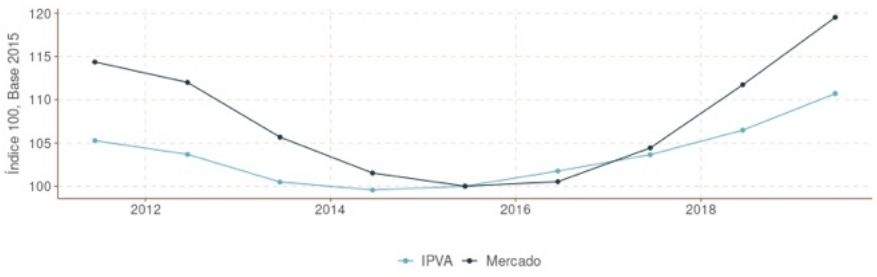
8.3.4 Comparativa con el IPVA

A continuación, se compara el IPVA del INE y los índices construidos a través de nuestra metodología. Es necesario advertir que aunque las magnitudes sobre las que se construyen son muy parecidas, existen diferencias entre ambos métodos:

- El IPV utiliza el precio registrado por el MITMA, que incluye solo rentas declaradas en el modelo 100, mientras que nuestro índice también incluye otro tipo de rentas, como las rentas sociales. Además el IPV exige que la muestra exista en el año en curso y en el anterior.
- El IPV utiliza un índice de Laspeyres encadenado y el nuestro es un índice de Fisher.
- Las cantidades usadas en el IPVA es un factor proporcional al gasto en euros del estrato, y el índice de mercado usa los metros cuadrados útiles del estrato. Además el IPV fija las cantidades para dos periodos.
- Aunque esta cuestión no debería ser fuente de grandes diferencias, los precios utilizados por el IPVA son en euros/metro cuadrado construido y nuestro índice en euros/metro cuadrado útil.

El IPVA muestra, en la Figura 8.37, cierto adelante con respecto al índice de mercado, además de una mayor divergencia en el primer y último año de la serie.

Figura 8.37. Comparativa indice de mercado e IPVA



Fuente: elaboración propia.

En la Tabla 8.14 se observa en los años más alejados de la base, los valores son mayores.

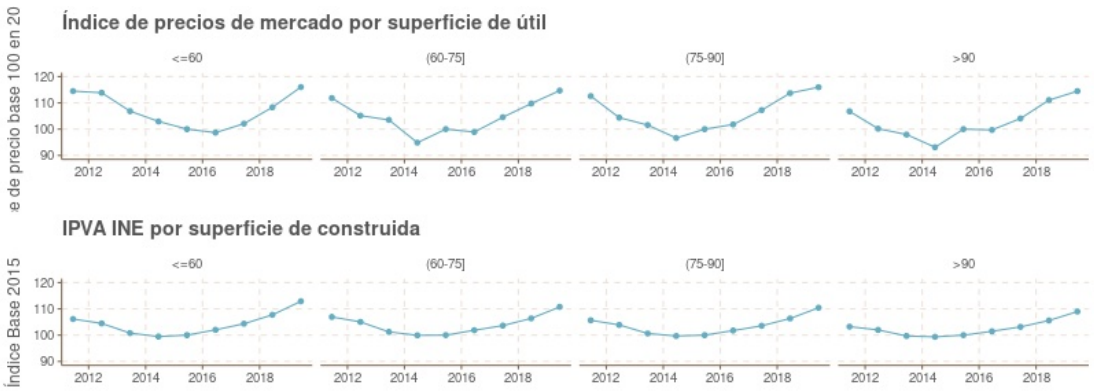
Tabla 8.14. Comparativa entre IPVA e Índice de Mercado

	2011	2012	2013	2014	2015	2016	2017	2018	2019
IPVA	105,3	103,7	100,5	99,6	100,0	101,8	103,7	106,5	110,7
Índice de Mercado	114,4	112,0	105,7	101,5	100,0	100,5	104,4	111,7	119,5
Diferencia	-9,1	-8,3	-5,2	-2,0	0,0	1,2	-0,8	-5,2	-8,8

Fuente: elaboración propia

En la Figura 8.38 se compara el índice desglosado por unidad de superficie para 4 rangos de precios[13], observándose como en términos de tendencia los dos índices muestran un comportamiento similar, aunque con un rango de variación más amplio en el caso de nuestro índice, sobre todo en el año 2015. En el caso de los inmuebles menores a 60 m², el IPVA muestra una tendencia creciente a partir del año 2015, mientras que en el índice de alquiler esto sucede a partir del 2017, producto de las diferencias en la composición muestral de las series.

Figura 8.38. Comparativa indice de mercado por superficie útil e IPVA por superficie construida

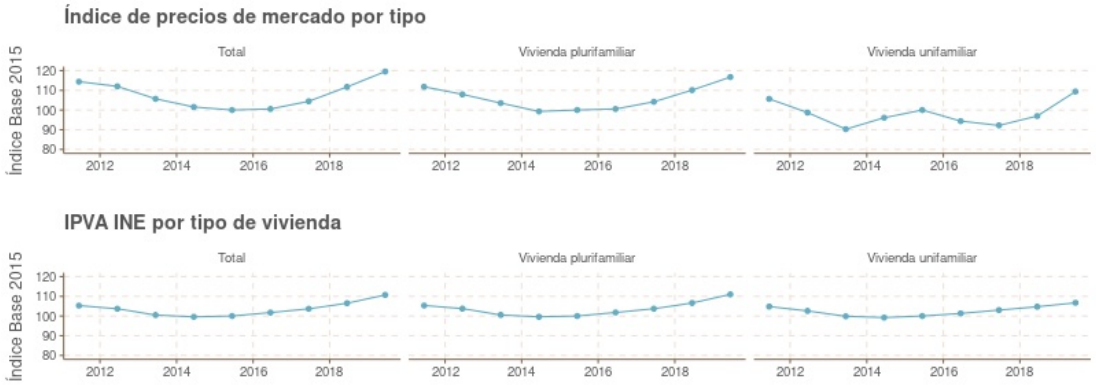


Fuente: elaboración propia.

Para el desglose por tipo de vivienda (Figura 8.39), se aprecia como existe un adelanto de los incrementos en los precios en el IPVA, con respecto a nuestro índice. Se puede atribuir al efecto de la vivienda unifamiliar, el cual muestra un comportamiento casi plano (y con alguna irregularidad en 2013) entre 2013 y 2016; en cambio el IPVA muestra un ligero crecimiento a partir del año 2014, de nuevo, motivado por las diferencias de la composición de la muestra en cada caso.

[13]Para el caso del IPV se ha creado la categoría “>90” como una media ponderada de los 3 rangos originales “(90- 120)”, “(121-150)” y “>150”.

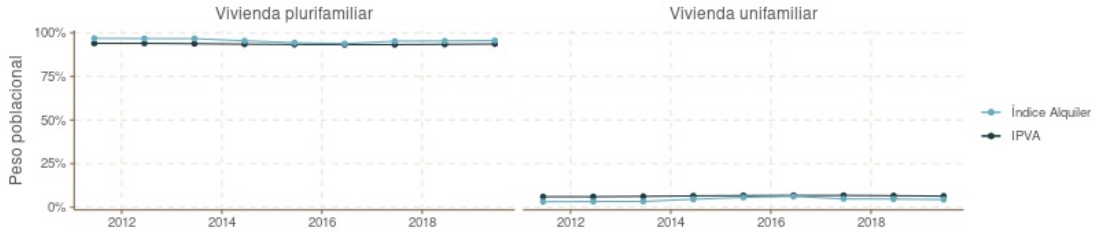
Figura 8.39. Comparativa indice de mercado por superficie útil e IPVA por tipo



Fuente: elaboración propia.

En la Figura 8.40, se observa que los pesos utilizados en ambas muestras son prácticamente coincidentes, como ambas proceden de la EPF, confirmaría la coherencia del proceso de calibración de poblaciones.

Figura 8.40. Evolución pesos del índice



Fuente: elaboración propia.

8.3.5 Capacidad predictiva del índice

Para evaluar la capacidad predictiva de los índices se plantea un modelo simple de proyección (*forecasting*) del precio del alquiler, basado en un modelo autorregresivo de orden 1 (véase Hyndman y Athanasopoulos (2018)). Puesto que se cuenta con dos series de precios, se decide compararlo con una aproximación multivariante. La muestra se compone de las series de variaciones para los años 2012 a 2019, y debido al reducido número de datos de las series, se utilizará una estrategia de remuestreo para estimar la precisión del modelo[14]. Esta aproximación asume el supuesto, ampliamente contrastado, de que existe inercia en los procesos de ajuste de los precios inmobiliarios, y que se refleja en la de los

[14]La estrategia de remuestreo, similar a la validación cruzada con K mezclas, se basa en realizar la validación 7 veces, cada una de ellas utiliza un año diferente como variable objetivo. Por tanto el primer muestreo usa el 2012 para validación y el resto de años (6) para modelar, en el segundo dejaría fuera el 2013, y así sucesivamente.

retornos[15] (Hwang y Quigley, 2010).

Debido el potencial predictivo de la serie de oferta se desarrollará un modelo autorregresivo con variables exógenas de tipo ARX(1,0)[16], siendo la variable endógena las diferencias del índice de alquiler, y la exógena las diferencias de la oferta. Formalmente, el modelo se especifica según la siguiente expresión:

$$Y_t^a = \alpha + \beta \cdot Y_{t-1}^a + \lambda \cdot X_{t-1}^o + \gamma \cdot X_t^o + \varepsilon_t \quad [8.9]$$

donde Y_t^a representa el vector de diferencias aritméticas del índice de alquiler $\Delta I_{e=s,t}^a$ para el estrato s entre los periodos t y $t-1$, y calculadas como: $I_{e=s,t}^a - I_{e=s,t-1}^a$. X_t^o es el vector equivalente para las series de oferta, y ε_t el término de error aleatorio. Los vectores X^o e Y^a tienen dimensión $[1, 6 \times 244]$, para el caso en el que la muestra se corresponde a los años 2013 a 2018 y la validación a 2019, definéndose analíticamente como:

$$\begin{bmatrix} \Delta I_{e=1,2013}^a \\ \dots \\ \Delta I_{e=1,2018}^a \\ \dots \\ \Delta I_{e=244,2013}^a \\ \dots \\ \Delta I_{e=244,2018}^a \end{bmatrix} = \alpha + \beta \begin{bmatrix} \Delta I_{e=1,2012}^a \\ \dots \\ \Delta I_{e=1,2017}^a \\ \dots \\ \Delta I_{e=244,2012}^a \\ \dots \\ \Delta I_{e=244,2017}^a \end{bmatrix} + \lambda \begin{bmatrix} \Delta I_{e=1,2012}^a \\ \dots \\ \Delta I_{e=1,2017}^a \\ \dots \\ \Delta I_{e=244,2012}^a \\ \dots \\ \Delta I_{e=244,2017}^a \end{bmatrix} + \gamma \begin{bmatrix} \Delta I_{e=1,2013}^o \\ \dots \\ \Delta I_{e=1,2018}^o \\ \dots \\ \Delta I_{e=244,2013}^o \\ \dots \\ \Delta I_{e=244,2018}^o \end{bmatrix} + \varepsilon_t \quad [8.10]$$

De forma equivalente, se construye un modelo autorregresivo AR(1) para evaluar el beneficio de utilizar la información de oferta:

$$Y_t^a = \alpha + \beta \; Y_{t-1}^a + \varepsilon_t \quad [8.11]$$

Es importante aclarar que los modelos autorregresivos requieren la estacionariedad[17] de las series. En este caso particular, al trabajar con series muy cortas, la capacidad predictiva de las pruebas habituales de detección de estacionariedad en series temporales está significativamente limitada. Según Schwert (1989)

[15] Los retornos inmobiliarios se refieren a los rendimientos financieros de una propiedad, como por ejemplo, el rendimiento bruto del alquiler, calculado como los ingresos anuales del alquiler divididos por el precio total del inmueble.

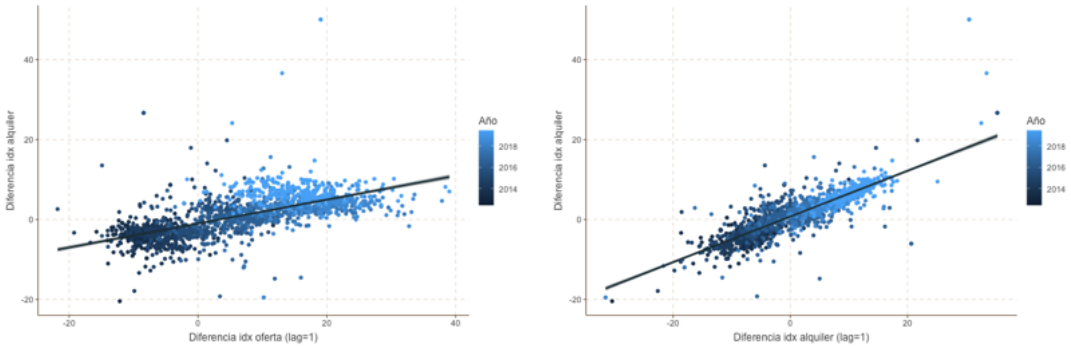
[16] Un modelo ARX es un tipo de modelo autorregresivo lineal con variables exógenas, de uso común en la proyección de series temporales. Se basa en capturar la relación de la variable dependiente en un periodo con datos de periodos anteriores, junto con el aporte de otras variables exógenas independientes, para más información véase Box _et al. (2016) o Chen y Tsay (1993).

[17] Una serie temporal es estacionaria cuando la media y la desviación típica son constantes en el tiempo, esta condición es necesaria para el desarrollo de ciertos modelos de proyección de series.

y DeJong *et al.* (1992), la potencia de las pruebas de raíz unitaria, como la de Dickey-Fuller, es baja en muestras pequeñas. Particularmente, Schwert menciona que la potencia de esta prueba tiende a ofrecer una escasa confianza cuando el tamaño de la muestra es menor a 50 observaciones.

A pesar de la incapacidad para testar la estacionariedad, se evalúan la correlación ρ entre las variables, obteniendo respectivamente: $\rho(\Delta I_t^a, \Delta I_{t-1}^a) = 0,88$, $\rho(\Delta I_t^a, \Delta I_{t-1}^o) = 0,73$ y $\rho(\Delta I_t^a, \Delta I_t^o) = 0,64$. De forma gráfica se puede confirmar la fortaleza de la correlación (Figura 8.41).

Figura 8.41. Relación diferencias de alquiler y oferta



Fuente: elaboración propia.

La Tabla 8.15 muestra una comparativa entre los dos modelos en términos de ajuste en R^2 y en errores medios absolutos y porcentuales, en general y desglosados por ámbitos geográficos. Los dos modelos tienen un alto R^2, ligeramente menor al 80%, que es sensiblemente mayor en el modelo ARX(1,0). En términos de error no hay grandes diferencias tampoco, aunque se observa que aunque el comportamiento general es mejor en el ARX(1,0), cuando se desglosa por zona se obtiene un mejor rendimiento en el modelo autorregresivo AR(1), lo que se puede atribuir a efectos de composición.

Tabla 8.15. Comparativa de modelos de forecasting de índices

	AR(1)			ARX(1,0)		
Ámbito geográfico	**R2**	**MAE**	**MAPE**	**R2**	**MAE**	**MAPE**
Madrid		1,589	38,20%		1,638	39,38%
Resto	78,09%	1,952	57,14%	78,89%	1,974	57,78%
Todas		1,078	31,55%		0,971	28,43%

Fuente: elaboración propia

Este modelo corrobora la adecuación de las series temporales para ser aplicadas

en un estudio que proyecte las variaciones futuras en el mercado inmobiliario. No obstante, el análisis podría extenderse hacia una perspectiva econométrica más exhaustiva, incorporando elementos asociados al comportamiento a largo plazo, como las transformaciones demográficas, las variaciones en los ingresos, los factores macroeconómicos y las tasas de interés[18], ajustándose a las particularidades de cada municipio. Este enfoque ha sido llevado a cabo por Taltavull[19] en un estudio que abarcó 275 municipios españoles con una población superior a 25.000 habitantes .

En el presente capítulo, se ha desarrollado un análisis exhaustivo de los índices de precios construidos, que ha permitido confirmar su calidad desde una perspectiva estructural y de mercado. Dicha verificación ha tomado en consideración su coherencia con índices de naturaleza similar, como el IPV desarrollado por el INE. Además, el modelo propuesto ofrece una herramienta de análisis dual sobre el mercado inmobiliario, al permitir evaluar simultáneamente los comportamientos temporales de oferta y alquiler, fundamental no solamente para el estudio de la evolución de dichos comportamientos, sino también, para la identificación de los factores subyacentes.

El estudio del aspecto geográfico revela diferencias notables en el comportamiento de los precios según la zona. Considerando los resultados obtenidos, se observa que las áreas densamente pobladas y urbanas de mayor nivel adquisitivo, y las viviendas con características constructivas y de instalaciones más apreciadas por los consumidores experimentaron un crecimiento más pronunciado en los precios de la vivienda, especialmente en el caso de la capital. Estos hallazgos parecen confirmar que las áreas más pobladas y urbanas de lujo de la Comunidad de Madrid han experimentado un comportamiento superior comparado con las áreas rurales y menos exclusivas, al menos en términos de recuperación de los precios de la vivienda. Lo cual confirma la hipótesis de que las áreas de alta demanda, muestran una recuperación rápida de los precios, mientras que en zonas con menor dinamismo inmobiliario, la recuperación es más lenta.

Por último, también se desprende de los resultados obtenidos el potencial de las series de números índice producidas para proyectar los comportamientos futuros, en el corto plazo, del mercado inmobiliario.

[18]Case y Shiller (2003); Ambrose *et al.* (2013); Hwang y Quigley (2010)
[19]-La Paz *et al.* (2021)

Anexo I. Conjunto de resultados

A continuación se indican una serie de recursos en Internet donde se pueden consultar los resultados del presente trabajo:

- Conjuntos de datos y gráficas de los índices de accesibilidad (capítulo 4): https://github.com/davidreyblanco/accessibility
- Gráficas de zonas antes y después del modelo hedónico final (capítulo 6): https://github.com/davidreyblanco/idx/tree/master/hedonic-final/unbias
- Gráficas detalladas de zonas antes y después del modelo hedónico final (capítulo 6): https://github.com/davidreyblanco/idx/tree/master/hedonic-final/unbias-fixed
- Aplicación interactiva para mostrar los índices de alquiler (capítulo 8): https://priceindexweb-mad-interactive.streamlit.app

Conclusiones

“Las reglas de la moralidad no son la conclusión de nuestra razón.”

— David Hume

La presente investigación ha logrado desarrollar satisfactoriamente un índice de precios de la vivienda en alquiler con los objetivos pretendidos: alto nivel de detalle temporal y geográfico, fiel reflejo del mercado mediante una alta consistencia con las fuentes oficiales y el uso de fuentes abiertas con métodos de aprendizaje estadístico.

Durante el transcurso de la misma, y para la consecución del objetivo, se han abordado de manera sistemática un conjunto de análisis que abarcan desde tener una perspectiva general del sector, hasta la definición concreta de los diferentes procesos metodológicos. A lo largo del proceso se han obtenido importantes conocimientos.

El mercado español de la vivienda presenta un comportamiento diferencial al resto de los países de la Unión Europea. En nuestro país, existe una proporción mayoritaria de hogares en propiedad desde finales del siglo pasado. Sin embargo, tras la última gran crisis inmobiliaria se ha comenzado a revertir esta situación, motivada, no solamente, por un cambio actitudinal en las preferencias entre comprar o alquilar, sino también por la restricción crediticia de los primeros años de la década de 2010, la cual ha abocado a los grupos socioeconómicos menos favorecidos a tomar el alquiler como única alternativa. En paralelo, la presión de la demanda ante una oferta apenas creciente, en las áreas más densamente pobladas, ha dado lugar a un incremento progresivo de los precios del alquiler.

1. Un factor importante en la escasez de oferta de vivienda fue consecuencia de la limitada construcción de vivienda nueva durante casi una década. Esto tiene su justificación en la explosión de la burbuja inmobiliaria, unida a un crecimiento de la población de más de dos millones de personas

(principalmente en las áreas metropolitanas). La restricción crediticia hipotecaria desde finales de los 2000 hasta mediados de la década del 2010, impulsó de forma notable la demanda de vivienda en alquiler. Ambos factores han forzado un nuevo equilibrio entre oferta y demanda, que ha producido un incremento sostenido de los precios desde el año 2015.

2. España es un mercado con una oferta de alquiler generalmente inelástica, particularmente en las grandes áreas metropolitanas. Dichas zonas, cuya demanda ha sido creciente en la última década, experimentaron un notable incremento de precios, alimentado por una oferta estática o decreciente, producto de la escasez de vivienda nueva y social. Lo cual ha producido grandes diferencias en los mercados en función de la geografía, siendo las capitales más pobladas las que han sufrido mayores subidas de precios.
3. El fuerte incremento de los precios afectó principalmente a los colectivos menos favorecidos, como los jóvenes o familias con menos recursos, por su casi imposibilidad de acceso al crédito. Lo cual es especialmente problemático en las grandes áreas metropolitanas del país, donde se concentra una gran demanda de vivienda junto con una oferta limitada. Dicha situación han contribuido a la degradación de las condiciones de vida y expulsión hacia zonas más económicas de los colectivos anteriores.
4. La escasez de vivienda social está vinculada a una regulación del mercado de alquiler en España que no han favorecido la construcción pública de este tipo de vivienda. Además, anteriormente y durante un periodo prolongado (previo al estallido de la última burbuja) se había favorecido el régimen en propiedad. Los cambios de los últimos 10 años se han centrado en actuar sobre otros aspectos, como la implantación de incentivos fiscales para arrendadores y arrendatarios, las políticas de limitación de precios en ciertas zonas, y el control de la actualización anual de las rentas a un máximo por área. Aunque estas medidas animan a los propietarios privados a poner en el mercado vivienda social, esta es insuficiente en número e incentivos.
5. Desde el punto de vista del fomento de la oferta, la normativa española ha asignado un rol principal al mercado privado como principal proveedor de vivienda en alquiler, tanto como promotor de vivienda social nueva como arrendador de los colectivos menos favorecidos. Con el fin de aliviar la carga sobre los propietarios individuales y fomentar la oferta, se otorgan incentivos fiscales a los propietarios y desincentivos o penalizaciones para los propietarios de viviendas desocupadas. Aún así, esta medida se ha mostrado insuficiente en impacto real y debería acompañarse una importante inversión en promoción de vivienda pública social.
6. El aspecto anterior es, en parte, consecuencia de una normativa que

favorece los alquileres plurianuales y de la inestabilidad regulatoria (cambios frecuentes de las políticas de vivienda en los ámbitos nacionales y regionales). La cual produce, consecuentemente, que un porcentaje de los propietarios decidan dejar este mercado, realimentando la escasez en la oferta. Por tanto, la normativa debería proteger la situación legal de los mismos, para evitar este efecto pernicioso.

7. Para evitar el fuerte incremento de las rentas en los núcleos urbanos más densamente poblados de los últimos años (Madrid, Barcelona, Málaga, por ejemplo), se han articulado varias acciones orientadas a controlar el precio máximo en las zonas cuyos precios están más tensionados (la relación precio medio del alquiler / ingresos es muy alta). A la vista de los resultados, estas actuaciones no han tenido el efecto esperado porque, entre otras cuestiones, desincentivan a parte de los propietarios para mantener sus inmuebles en mercado, acrecentando la tensión sobre la oferta.
8. La principal crítica a las políticas restrictivas de precios, denominadas de segunda generación, es que introducen distorsiones entre las relaciones de los mercados de compra y alquiler, ignora aspectos fundamentales en la formación de los precios y, en el largo plazo, desincentiva las inversiones del propietario en mantener correctamente la propiedad. Por otra parte, tanto en el mercado nacional, como en los países del entorno, ninguna de las normativas de este tipo ofrecen evidencias claras sobre su eficacia.
9. Dentro de la administración no existe consenso respecto a quién debe liderar el desarrollo de la política de vivienda, planteando la duda de si estas deben ser nacionales, o si las comunidades autónomas o los ayuntamientos deberían ser los últimos responsables de decidir las normas a implementar. Parece evidente que las políticas deben articularse desde las entidades locales dentro de un marco general de política de la vivienda estatal o regional. De forma que se ataje el problema secular de la ausencia de vivienda social, a través de actuar sobre la oferta mediante un plan ambicioso de desarrollo de vivienda de esta naturaleza.
10. En paralelo a los puntos anteriores, en términos microeconómicos, la asimetría de información entre los oferentes y los demandantes de vivienda conduce a una escasez artificial, producto de la falta de confianza en el arrendatario, que se traduce en una prima de riesgo ante la posibilidad de impagos. Por lo cual el desarrollo de herramientas que ofrezcan más solvencia del inquilino y la creación de financieros que ofrezcan garantías ante los propietarios, para las personas más desfavorecidas, ayudarían a reducir las primas de riesgo en las cuotas.

La contabilidad nacional es la referencia estadística y se acepta

comúnmente como fuente de las principales actividades de la economía. Sin embargo, presenta una serie de inconvenientes. En primer lugar, ofrece información muy agregada y con un retraso temporal importante para el sector inmobiliario. Por otra parte, el uso generalizado de Internet y aplicaciones móviles para la búsqueda de la vivienda ha permitido que dichas plataformas tecnológicas, sean una fuente de información relevante en el estudio de los fenómenos inmobiliarios. Sin embargo, el uso de este tipo de datos en la construcción de índices de precios de la vivienda plantea retos metodológicos producto de sus sesgos y la desigual calidad de la información.

11. Al contrario de lo que sucede en el sector inmobiliario de compraventa a nivel nacional, la administración no dispone de herramientas de información sobre los precios del alquiler adecuados para el estudio exhaustivo de las dinámicas del mercado. El último intento de la administración central al respecto fue la creación, en 2020, de un índice de precios del alquiler, dirigido por el Ministerio de Transportes y elaborado en colaboración con otros organismos públicos. A pesar de este esfuerzo, dicho índice no satisfacía completamente todos los requisitos deseables, como el de contar con datos recientes, detallados y desglosados geográficamente.
12. Aunque existen fuentes de información privadas con un alto grado de actualización y desglose (portales inmobiliarios), estas no contienen precios finales de rentas, sino datos de oferta. Además, solo recopilan el producto anunciado en sus plataformas, lo que impide obtener una visión íntegra del mercado, que analizadas de forma aislada conforman una imagen sesgada de los mercados inmobiliarios.
13. Al contrario que en el uso de las fuentes estadísticas tradicionales, los datos de portales inmobiliarios adolecen de problemas estructurales de calidad de la información (duplicados, valores extremos, por ejemplo). Además, la diversidad de fuentes, magnitudes y métodos de elaboración exigen la homogenización y consolidación de las mismas.
14. Los datos registrados en los portales de Internet representan una perspectiva sesgada de la oferta y provienen de registros introducidos por usuarios. Por tanto, es necesaria la aplicación de procesos de tratamiento previo de la información tales como: eliminar valores atípicos, corrección errores, imputación de valores ausentes y eliminación de registros duplicados.
15. Se requiere información muy detallada en todas las fuentes utilizadas, particularmente en lo relativo al aspecto geográfico, ya que la heterogeneidad espacial asociada al precio no siempre puede explicarse mediante los atributos de la vivienda o tipologías de zona.

16. Los cambios metodológicos en las fuentes oficiales a lo largo del tiempo influyen de manera significativa en la estabilidad temporal de las medidas. En el presente caso, ha sido necesario aplicar un proceso de suavizado exponencial en el dato de la EPF para solventar este problema.
17. El empleo de múltiples fuentes con grandes volúmenes de datos (decenas o cientos de millones de registros) requiere la aplicación tecnología de procesos de datos eficiente y a gran escala, a través de plataformas tecnológicas en "la nube".

Los métodos de correspondencia estadística (*statistical matching*) son eficaces para replicar la estructura poblacional original de la muestra a través de los elevadores muestrales y mitigar el efecto de los sesgos de no respuesta (indisponibilidad de información para uno o varios estratos de la muestra), de composición, variables omitidas, de infrarrepresentación y sobrerrepresentación, habituales en los datos procedentes de portales inmobiliarios de Internet. La correspondencia estadística se aplica tanto en el muestreo como en la función que traduce los precios de oferta en precios de mercado.

18. Los sesgos muestrales comunes en los datos de oferta (obtenidos de Idealista) incluyen el sesgo de no respuesta, el sesgo de composición, así como el sesgo de infra y sobrerrepresentación. Todos estos sesgos se deben a factores exógenos al mercado y distorsionan la percepción sobre el mismo. Algunos de ellos proceden, por ejemplo, de la cuota de mercado del portal, una mayor orientación hacia las agencias en detrimento de los particulares, o al nivel de adopción del servicio de los agentes inmobiliarios en una zona.
19. El proceso de calibración de poblaciones permite ajustar de forma apropiada los pesos poblacionales y enlazar fuentes diversas en una estratificación compleja, sin embargo requiere de variables auxiliares que vinculen los conjuntos de datos a conectar estadísticamente.
20. Cuando no se dispone alguna de las características en los dos conjuntos es posible estimarla mediante modelos. En nuestro caso, el modelo utilizado ha dependido de la naturaleza de la variable: año de construcción (a través de un proceso geográfico de imputación del año de construcción catastral), la superficie útil de la vivienda, o los ingresos familiares (a través de un modelo de tipo *Random Forests*).
21. El procedimiento de calibración en múltiples etapas resulta ventajoso para realizar la correspondencia cuando las variables comunes no están totalmente disponibles en todos los registros, y no es posible imputar la información a través de modelos. Para el presente caso, las magnitudes de

precios de oferta y reales de alquiler se vinculan mediante la combinación de calibración y modelos de regresión hedónica.

22. Sin embargo, la calibración presenta varias limitaciones que afectan a su eficacia, como se ha podido constatar en nuestro caso de aplicación:
    a. La estabilidad de las precios para un segmento heterogéneo, escasamente representado y con pesos irregulares a lo largo del tiempo, como sucede con las viviendas unifamiliares, es generalmente baja. Incluso en el caso en el que el nivel de ajuste en los modelos hedónicos es alto, la dispersión es mucho mayor comparada con estratos más representativos.
    b. Ante viviendas más complejas, como sucede con las viviendas unifamiliares, y considerando un menor tamaño muestral, la irregularidad en los resultados de los índices de precios de viviendas puede deberse a la omisión o la imprecisión de variables explicativas clave. Por ejemplo, la limitación máxima de superficie útil en la EFF a m^2 dificulta la especificación de los modelos de valoración de viviendas con superficies amplias.
    c. La incorporación de zonas exactas resulta esencial para el correcto desempeño de los modelos, dada la presencia de heterogeneidad espacial en los precios. Cuando la información no está disponible en la fuente, como es el caso de la base de datos de la EPF, se requiere un nuevo proceso de reponderación basado en fuentes detalladas, como se realizó con los datos de precio del MITMA.
    d. Para abordar el problema espacial, se plantean diferentes alternativas. La primera consiste en efectuar un ajuste de sesgo posterior utilizando una fuente auxiliar (como en el caso del MITMA), que puede realizarse por medio de ratios o mediante un modelo lineal o de árboles. La segunda opción sería reemplazar la EPF por una encuesta que abarque todas las áreas de estudio y que permita construir el modelo de conversión considerando la ubicación del inmueble.
    e. La limitada granularidad de los microdatos de la EPF introduce una incertidumbre importante en el modelo de correspondencia de precios de oferta y mercado (correspondencia estadística de los precios). Este fenómeno conduce a errores en la estimación de los mismos en algunos estratos, los cuales requieren un ajuste posterior.
    f. Generalizando, aunque la correspondencia estadística aporta una solución que, de forma general, permite replicar la estructura muestral y la relación entre magnitudes de precios de oferta y reales, presenta desafíos en términos de representatividad, precisión y disponibilidad de variables y datos espaciales.

La construcción de herramientas de supervisión y control de los mercados inmobiliarios plantea retos como consecuencia de la naturaleza altamente heterogénea de la vivienda. Lo cual implica que la construcción de índices de precios de la vivienda sea una disciplina compleja y que invulcra numerosos aspectos (geografía, características constructivas o estado de conservación, por ejemplo). Para abordar la cuestión, la metodología se plantea en dos planos: el de los precios, donde se estudian los factores que los determinan; y el temporal, que se estudia la evolución temporal de los valores.

23. En décadas recientes, los organismos encargados de establecer mecanismos macroprudenciales han exigido a las agencias estadísticas nacionales un control sobre los índices de la vivienda, lo cual ha derivado en la homogeneización de estas técnicas estadísticas. Aún así, no existe un consenso metodológico respecto a las variables a considerar para construir los modelos, ya que las realidades constructivas y urbanísticas de cada región determinan la idoneidad de las características a emplear.
24. Entre de las técnicas más idóneas y, por tanto, recomendadas por las agencias estadísticas, se encuentran aquellas basadas en modelos hedónicos, y que permiten la descomposición de los factores fundamentales del precio. Sin embargo, dada su complejidad de cálculo, un gran número de los índices de precios desarrollados por las agencias estadísticas estatales emplean modelos de ajuste mixto de medianas.
25. Los modelos hedónicos paramétricos basados en regresiones lineales poseen la ventaja de ser altamente interpretables, pero conllevan numerosas dificultades en su especificación funcional. Como alternativa a los métodos lineales, recientemente se han popularizado métodos hedónicos que emplean aprendizaje estadístico que resuelven, en gran medida, las debilidades de los primeros, pero a costa de una menor interpretabilidad en su funcionamiento.
26. Los métodos de aprendizaje estadístico estiman los coeficientes de cada una de las covariables del modelo, lo que conduce a que los índices de precios hedónicos, construidos sobre los modelos, sean necesariamente los denominados de doble imputación. Esta aproximación aporta la ventaja de ser más eficaz ante el sesgo de omisión de variables, que es frecuente en este los modelos de valoración de la vivienda, al no disponer de toda la información del bien inmobiliario (principalmente por la complejidad y naturaleza única del bien).

El modelo hedónico es un mecanismo eficaz que permite descomponer el precio de la vivienda en base a sus determinantes fundamentales. En este

caso, dado el gran número de factores involucrado, la descomposición del proceso en distintos aspectos (funcional, geográfico y tipología) y su consolidación en un modelo ensamblado, ofrece un buen balance entre complejidad del modelo y precisión. El uso de Random Forests ensamblados como método de regresión permite solucionar algunos de los retos de la modelización hedónica tradicional (multicolinealidad, heterocedastidad, no linealidades, entre otros), ofreciendo buenos niveles de reducción de la varianza y sesgo.

27. No existe un conjunto canónico de variables a utilizar, ya que la idoneidad está marcada por el tipo de vivienda y zona geográfica. No obstante y para evitar este inconveniente, se han tomado las categorías de variables más utilizadas en la academia y la industria, como son: elementos constructivos, morfología del edificio y de la vivienda, equipamiento, acceso a servicios cercanos, características de la zona y aspectos del mercado inmobiliario cercano.
28. *Random Forests* permite tratar una gran cantidad de variables sin incurrir en los problemas de la regresión lineal generalizada, controlando eficazmente los fenómenos de heterocedasticidad, multicolinealidad, las no linealidades y las interacciones complejas entre variables. Además, los algoritmos de árboles ofrecen muy buenos resultados en términos de ajuste con muy bajo sesgo. Dos ventajas adicionales de este método es su baja tendencia al sobreajuste y al hecho de ofrecer un mayor número de estadísticos en el proceso de entrenamiento (a través de las métricas sobre la muestra OOB).
29. Este modelo es menos sensible a las distorsiones introducidas por valores extremos que la regresión ordinaria. No obstante, tiende a excluir los valores menos usuales, y por tanto, la estimación de los mismos es poco fiable. En nuestro caso particular, se observa como el modelo ofrece peores resultados en precios por metro cuadrado anormalmente altos, mientras que no son problemáticos los casos extremos inferiormente.
30. El análisis experimental sugiere que la especialización de los modelos ofrece mejores resultados en escenarios en los que interviene un gran número de variables. La combinación de todos estos modelos en un ensamblaje final permite lograr un ajuste más equilibrado en todos los estratos, como se evidencia por un aumento en el coeficiente de determinación (R^2) en todas las zonas.
31. El ensamblado permite separar la capacidad de capturar la varianza desde distintos ángulos (por los modelos individuales) y corregir el sesgo de cada uno de ellos en el proceso de combinación del resultado final. Un número excesivo de modelos a ensamblar también podría resultar problemático

por tanto se decide especializar en dos módulos: atributos de la vivienda y localización. Para lograr la unión eficaz los módulos se comparten atributos zonales (tipo de zona) y morfológicos (superficie y tipo de construcción).

32. Los modelos construidos mediante ensamblado no están exentos de problemas. Algunos de los inconvenientes encontrados en su aplicación son su complejidad, la generación de un gran número de modelos y su capacidad reducida para ser interpretados y analizados en detalle.
33. Dado que los niveles de precios por metro cuadrado, las características y el nivel de soporte muestral de las viviendas unifamiliares y las plurifamiliares son muy diferentes, el uso de un modelo por cada tipo aporta un mecanismo de control adicional en la estimación del valor de la vivienda.
34. La selección del precio por superficie como variable objetivo reduce los errores de estimación en los valores muy altos. Aunque, desde el punto de vista inmobiliario, puede producir problemas de especificación de los modelos en las viviendas unifamiliares. En dichas propiedades, la superficie de la parcela es un elemento determinante en el valor final, pero no lo es en el precio la vivienda por unidad de superficie construida.

La vivienda es un bien inusual en tres aspectos: heterogeneidad, durabilidad e inmovilidad. Este último factor señala a la localización como un elemento clave en la construcción de modelos de precios hedónicos. No obstante, los principales retos al introducir la ubicación en este tipo de métodos son: la inexistencia de una forma canónica de especificarla, la complejidad derivada de la heterogeneidad espacial y la dificultad de cálculo de los modelos de regresión geográfica.

35. No se puede desligar el análisis espacial del modelado hedónico. El adecuado funcionamiento de los modelos hedónicos está condicionado por los efectos en el precio a lo largo del espacio, ya que las relaciones entre los predictores y el precio pueden variar en función de la zona (heterogeneidad espacial). Por tanto, un buen predictor en una zona puede no ser tan efectivo en otra, lo que invalida ciertas condiciones fundamentales de los modelos de regresión si no se incorporan variables que especifiquen correctamente la interacción con la geografía. Asimismo, es importante mencionar que muchos de los determinantes de los precios vinculados a la localización son desconocidos (por ejemplo, cuestiones actitudinales como las preferencias de los inquilinos por ciertas zonas que son muy difíciles de incorporar como atributos de los modelos).
36. Para incorporar de forma eficaz la influencia geográfica en los modelos manteniendo el principio de valoración hedónica, es posible especificarla

sobre el nivel de utilidad asociado a la localización. Esta magnitud se define como el grado de acceso a distintos servicios, medidos en tiempo de desplazamiento. Desde un punto de vista numérico se expresan a través de unos índices de accesibilidad en función de los elementos de posible interés del inquilino y que se encuentran cerca del inmueble.

37. La definición de variables de accesibilidad es compleja desde un punto de vista metodológico y de cálculo. Por este motivo, en muchos casos, la literatura aporta soluciones parciales (uso de distancias euclidianas en lugar de tiempos de transporte) y arbitrarias (factores seleccionados por criterio del investigador sin mayor justificación). Estas aproximaciones parciales pueden no incorporar completamente los factores que influyen sobre el precio.
38. La síntesis de la accesibilidad en pocas variables (existe una gran variedad de aspectos), y que sean significativas para el problema a resolver no es un proceso trivial. Nuestra propuesta de índices gravitatorios basados sobre isodistancias en tiempos de desplazamiento, y seleccionadas de forma automática por su capacidad de reducción de la autocorrelación espacial, demuestra ser un método de especificación general, eficiente y sencillo.
39. El uso de modelos de aprendizaje estadístico de árboles complejos y la accesibilidad permite que los propios modelos incorporen la diversidad de comportamientos a lo largo del espacio. Además, el amplio rango de dimensiones de accesibilidad construidas, permite incorporar la influencia desde un gran número de aspectos. Se demuestra, estadísticamente, que estas variables de accesibilidad reducen la dependencia espacial de los residuos de los modelos y, en el caso de *Random Forests*, prácticamente los eliminan.
40. Dado que las características de utilidad pueden estar correlacionadas entre si es conveniente aplicar un proceso de factorización (por ejemplo, las medidas de utilidad de servicios de restauración y turismo suelen ser bastante similares en los centros de ciudades turísticas).
41. Cuando se desarrollan variables de localización sintéticas (factorizadas), a menudo resulta difícil estimar su grado de contribución. En nuestro caso, el método basado en factorización permite ofrecer una semántica al campo por dos razones: 1) el proceso de construcción selecciona las variables que potencialmente más contribuirían a explicar el precio, y 2) porque el proceso de análisis de componentes principales describe las contribuciones en los aspectos de la variable original y el predictor sintético.
42. A pesar de que el método ofrece un buen rendimiento predictivo, aún quedan ciertas cuestiones pendientes que deben ser abordadas en investigaciones

futuras. Tales como la extensión del método para que sea multivariante, y la integración de otras técnicas de econometría espacial en el proceso de selección heurística de medidas de utilidad.

El modelo hedónico final, que estima el precio de mercado a través de correspondencia estadística con el precio de oferta, requiere incorporar información zonal a través de fuentes auxiliares. Como la unión se realiza anualmente y el número de modelos a desarrollar es muy amplio, podrían surgir inconsistencias de los modelos a lo largo del tiempo. El método utilizado demuestra funcionar adecuadamente en este último aspecto, ya que los modelos para un año dado predicen con bastante exactitud los precios para el año siguiente.

43. Al igual que en los modelos de regresión convencionales, aquellos segmentos con menor representación exhiben un mayor grado de error en la predicción, como ocurre con las viviendas unifamiliares.
44. Los modelos ensamblados de árboles de regresión tienen una gran capacidad para capturar la varianza y producir modelos insesgados, pero presentan cierta rigidez estructural en términos de generalización. Lo anterior se deriva de que se asigna a nodo cada hoja del árbol el comportamiento promedio de una serie de observaciones, lo que dificulta la generalización de comportamientos en espacios de datos dispersos, escasos y altamente heterogéneos (como es frecuente en datos inmobiliarios). Alternativamente, se podrían proponer métodos más parsimoniosos que capturen la interacción espacial de forma general, y emplear esta estimación como entrada otro modelo (de tipo: bayesiano multinivel, basado en aprendizaje profundo, GAM, red neuronal o espacio-temporal).
45. A pesar del gran número de fuentes utilizadas, la información disponible es limitada en algunos estratos, particularmente en vivienda unifamiliar; en consecuencia, la exactitud del modelo podría mejorarse con la introducción de atributos específicos que representan sus singularidades, como fotografías del inmueble, textos de anuncios o incluso con imagen satélite de la finca. Alguno de ellos ya están presentes en los portales inmobiliarios, pero requieren un proceso avanzado de la información (visión por computador o minería de textos, por ejemplo).
46. El planteamiento de un modelo hedónico para cada año (impuesto por la restricción de datos en el proceso de correspondencia estadística) no garantiza la continuidad temporal de los valores estimados. Lo cual se refleja en variaciones abruptas de los precios estimados entre el mes de diciembre y el enero del año siguiente (también por la tendencia del árbol

a sobrerrepresentar el comportamiento anual medio, en detrimento de las diferencias mensuales). Para solucionar lo anterior, se podría plantear modelo único para todos lo años, que a su vez presentaría los siguientes inconvenientes: 1) al incluir datos de un nuevo periodo, se podrían re-estimar datos de periodos anteriores en base a datos futuros, 2) el volumen de registros es muy elevado, y por tanto, se requerirían elementos más rigurosos para tratar la información.

47. Las fuentes oficiales usadas para la calibración de precios (EPF o censo) no cuentan con datos zonales exactos por tanto la correspondencia estadística en precios ha requerido el uso de una fuente de precios a nivel de zona exacta (datos de precios de alquiler de MITMA). Esto indica que las estadísticas agregadas por criterios funcionales o demográficos (zona urbana media, por ejemplo) combinan realidades de precios muy diferentes y su significancia es relativa.
48. La ausencia del control de zona exacta, descrita en el punto anterior, genera importantes problemas en la consistencia temporal de mercado en las series de precios (cambios de tendencia inesperados en las series), principalmente debidos a efectos de composición.
49. La correspondencia entre los precios vincula de manera muy eficiente los valores de oferta con sus equivalentes de alquiler, a tenor de su alto coeficiente de correlación a lo largo del tiempo (particularmente en vivienda plurifamiliar). Además, se observa una fuerte correlación cruzada entre las magnitudes del mercado y la oferta, con un retraso temporal, con el primero respondiendo a los cambios de la oferta.

La construcción del aspecto temporal requiere el desarrollo del modelo para todos los periodos de la muestra. Sin embargo, es habitual que los datos disponibles no se encuentren en la frecuencia temporal más desagregada y, en particular, en esta investigación buena parte del dato público disponible es anual. Un proceso automático que estima el método de desagregación para cada zona, basado en la verosimilitud de una serie estructuralmente óptima, resulta eficiente para generar series de alta frecuencia de mayor calidad.

50. La construcción de un índice de muy alta frecuencia (mensual, semanal o incluso diaria) que tenga en cuenta fuentes públicas requerirá generalmente de un proceso de integración temporal, principalmente porque muchas de las últimas se entregan anualmente o con retraso.
51. Se aprecian dos fenómenos en la integración de múltiples frecuencias de datos: 1) fuertes discontinuidades entre los meses finales e iniciales del

siguiente año; y 2) inconsistencia entre la media de las series mensuales generadas al aplicar los modelos directamente sobre el dato de oferta. Históricamente, estas cuestiones se solventan a través de diversos métodos de desagregación temporal que, por otra parte, introduce la dificultad de decidir qué métodos utilizar.

52. La desagregación temporal en un contexto de gran número de series y alta heterogeneidad entre ellas (por ser mercados inmobiliarios muy distintos), requiere un tratamiento particular de cada serie y es automatizable:
    a. Existe un vasto número de métodos a aplicar, siendo el más apropiado aquel que depende principalmente de la naturaleza de la serie y de la calidad de las series originales y auxiliares.
    b. Un enfoque de selección de parámetros y método basado en la calidad de las series generadas, como en nuestro caso con el método propuesto de máxima verosimilitud, ofrece resultados sólidos, pero el resultado también está fuertemente influenciado por la calidad de las series indicadoras utilizadas para estimar la serie final de alta frecuencia.
    c. El criterio de verosimilitud ofrece una consistencia con el criterio experto, al sustentarse sobre los mismos criterios que toma un especialista para seleccionar un método sobre otro, logrando un mayor nivel de credibilidad del método de decisión.
    d. Los métodos de selección multicriterio de verosimilitud resultan fácilmente interpretables y pueden mejorarse mediante la adición de nuevos criterios al proceso de selección. Particularmente, en nuestro caso, se plantean tres áreas de mejora futuras: la introducción de nuevos criterios, la desagregación multivariante y el uso de funciones de densidad no paramétricas para la estimación de la verosimilitud.
    e. Sin embargo, el modelado hedónico determina en parte la calidad del proceso de desagregación, puesto que las series indicadoras se estiman a través de precios hedónicos de la oferta. A este respecto, es clave la forma en la que se especifica el aspecto temporal en los modelos y, particularmente, en los modelos ensamblado con alta dimensionalidad. En la presente investigación, se observa que al haberse incorporado la componente tiempo mediante variables dicotómicas *dummy* por mes, las mismas entran en competencia con otras variables altamente explicativas, como son la zona, la superficie o el estado de conservación, lo cual puede dar lugar una infravaloración de las contribuciones de las variables *dummy* de tiempo.

Un índice de tipo encadenado superlativo basado en doble imputación hedónica permite mitigar sesgos producidas por variables omitidas.

El método de encadenamiento permite realizar ajustes dinámicos para tener en cuenta los cambios en la composición de la muestra a lo largo del tiempo. Este enfoque común en los mercados inmobiliarios, ya que estos están influenciados por múltiples factores de carácter dinámico, tales como los demográficos, financieros, urbanísticos, regulatorios o comportamentales de los agentes.

53. El índice construido garantiza la consistencia con datos recogidos por fuentes oficiales a través de la calibración y la bondad del ajuste de los modelos hedónicos. Como el nivel de detalle de las fuentes es limitado, sería deseable, además, poder contrastar los resultados desglosados con otra información objetiva (por ejemplo paneles o encuestas de control).
54. El detallado desglose funcional y geográfico facilita un análisis profundo en la evolución de los precios. Como se ha observado, las diversas zonas exhiben distintas velocidades de recuperación en función del nivel de demanda particular. Por otra parte, el análisis conjunto de los índices de precios de oferta y alquiler constituye una excelente herramienta para evaluar dinámicas del mercado, como las tensiones en los precios.
55. La falta de soporte en ciertos estratos introduce ruido en los índices, como sucede en el caso de las viviendas unifamiliares. Este fenómeno aporta inestabilidad a las series desagregadas, cuyo efecto no se puede mitigar completamente mediante el uso de índices de precios encadenados. No obstante, dado el reducido peso de dichos estratos, los problemas no afloran significativamente en los índices más agrupados.
56. Los índices de precios demuestran un comportamiento similar, en general, al IPVA experimental del INE, aunque se detectan ligeras discrepancias en la escala, que se pueden atribuir a las diferencias en los métodos y la estructura poblacional de ambos enfoques.
57. La evolución de los precios de oferta representa un indicador adelantado de la coyuntura de los precios de alquiler. Además, el grado de correlación entre ambos precios es muy elevado, especialmente para los inmuebles de tipo plurifamiliar, lo que confirma la adecuación de esta fuente para construir índices de precios, una vez aplicadas las ponderaciones sobre un diseño muestral que permite corregir los sesgos poblacionales.
58. En el índice se pueden identificar las diferentes velocidades, en las fases de caída y recuperación, vinculadas a los diversos niveles de liquidez del mercado inmobiliario. Por ejemplo, las zonas de Madrid capital y coronas metropolitanas inmediatas son las primeras en caer y recuperarse, mientras que las zonas rurales o de menor densidad poblacional presentan una recuperación más lenta. Esta misma distinción se puede realizar en función

de la excentricidad de dichas caídas, siendo las caídas más pronunciadas en los mercados ilíquidos.

En el análisis experimental sobre un índice de precios construido para la Comunidad de Madrid en el periodo 2011 - 2019, se observa que las áreas densamente pobladas y urbanas de mayor nivel adquisitivo, y las viviendas con características constructivas y de instalaciones más lujosas experimentaron un crecimiento más pronunciado en los precios de la vivienda. Estos resultados parecen confirmar la hipótesis inicial de que las áreas más pobladas y urbanas de lujo de la Comunidad experimentaron un comportamiento mejor comparado con las áreas rurales y menos exclusivas. Estas diferencias se pueden relacionar con el mayor o menor dinamismo de dichos mercados.

59. El índice de precios generales muestra una tendencia decreciente en el alquiler desde el año base hasta 2016, seguido de un crecimiento sostenido hasta 2019. Las fases de crecimiento y decrecimiento varían ligeramente por zona, aunque en todas ellas se aprecian una fase de caída y una de recuperación. Estas desigualdades se manifiestan como una caída anticipada en la zona central y más poblada, siendo además la primera en recuperarse. En cambio, las zonas periféricas y rurales tuvieron una recuperación más lenta y valores inferiores en 2019 respecto al año base 2011.
60. El análisis zonal también revela que la oferta de viviendas actúa como un indicador adelantado del precio del mercado, aunque con diferentes escalas de valores, y este diferente nivel de intensidad está relacionado con una mayor liquidez de la zona. En este sentido, la capital, Madrid, es la primera zona en recuperarse y de las últimas en depreciarse debido a la alta demanda que presenta, independientemente del ciclo económico. Por el contrario, las zonas con menor poder adquisitivo tardan más tiempo en recuperarse y además, dado a su menor interés inmobiliario, no muestran caídas ni subidas de precios exageradas.
61. Las diferencias en el comportamiento de los precios entre la capital (Madrid) y el resto de la provincia, están motivadas por la mayor demanda del área urbana y, por otra, porque en las zonas de alto poder adquisitivo el precio se mantiene estable ante cambios en el ciclo económico o inmobiliario, por su mayor elasticidad en los precios ante las condiciones económicas.
62. El análisis basado en características constructivas revela que las viviendas más antiguas no resistieron la crisis tan bien como las viviendas más nuevas, aunque experimentaron crecimientos similares durante la fase de

recuperación a partir del año 2016.

63. En cuanto a la segmentación por superficie útil, las viviendas de mayor tamaño mostraron un mejor comportamiento en comparación con las viviendas más pequeñas, que experimentaron caídas más pronunciadas durante el periodo 2011-2016.
64. El análisis por tipo de vivienda y tipo de edificio muestra una mayor estabilidad y crecimiento en precios para viviendas unifamiliares y más exclusivas en el resto de la Comunidad, especialmente en áreas densamente pobladas y urbanizaciones metropolitanas de alta gama.
65. Las áreas rurales exhibieron los descensos más pronunciados con tasas de crecimiento más bajas, en comparación con las zonas intermedias y densamente pobladas, lo cual puede atribuirse a la mayor presión de la demanda en las zonas urbanas.
66. La presencia de ciertas instalaciones en la vivienda, como el aparcamiento y la piscina, mostró una diferencia significativa en la evolución de los precios en función de las mismas, especialmente en el caso del resto de la Comunidad. Por contra, otros equipamientos que intuitivamente se relacionan con una mayor calidad del inmueble, como el aire acondicionado, no mostraron diferencias significativas en la evolución de los precios.
67. Se puede trazar una relación entre la desigualdad zonal y su composición socioeconómica. En municipios de alto poder adquisitivo, por ejemplo la zona noroeste del eje de la A6 (Pozuelo de Alarcón, Majadahoda y Las Rozas), se presenta una disminución inicial en los precios hasta 2013, con una meseta sin crecimiento y una vuelta al crecimiento a partir de 2018. En contraste, en zonas metropolitanas de poder adquisitivo medio y bajo, se percibe una caída progresiva de precios hasta 2015, seguida de una recuperación a partir de 2016. En este contexto, la transferencia de precios de oferta al alquiler es más lenta en comparación con municipios de mayores ingresos. Se percibe también una diferencia en la intensidad de las caídas y subidas según el poder adquisitivo, con transiciones más suaves en las zonas más acomodadas.
68. Los precios de oferta y alquiler presentan comportamientos homogéneos a lo largo del tiempo en todas las áreas geográfica, aunque con discrepancias en algunas zonas puntuales como la Sierra Norte y Oeste, que experimentan caídas en oferta de menor intensidad que en alquiler.

El índice de precios construido ofrece una herramienta de análisis en profundidad del mercado inmobiliario de alquiler, por múltiples motivos: 1) alta desagregación temporal, zonal y funcional; 2) incorpora las perspectivas de oferta y mercado; y 3) mantiene la coherencia con las cifras generales del mercado publicadas por los organismos oficiales.

69. La capacidad de segmentar la información de precios (tanto en valor absoluto como en número índice), permite el análisis en profundidad de los fenómenos inmobiliarios en un mercado complejo como es el de la vivienda desde tres aspectos en paralelo: zonal, funcional y temporal. Este potencial de análisis es esencial en el desarrollo de una política de la vivienda eficaz por parte de las distintas entidades nacionales, autonómicas y locales.
70. Esta metodología es replicable y extensible con datos locales y mucho más granulares, por tanto, se podrían desarrollar nuevos índices con mayor desagregación y con datos específicos para municipios o comarcas.
71. Por construcción, este índice mantiene la coherencia entre cifras generales publicadas por los organismos oficiales (INE o MITMA). Lo que permite su mejor combinación con fuentes de información públicas y fomenta su potencial adopción por las entidades públicas encargadas de la política de la vivienda.
72. El modelo que vincula el precio de oferta y el de mercado, permite analizar en profundidad, entre otras cuestiones, las relaciones funcionales entre oferta y demanda, el poder de negociación de propietarios y arrendatarios. En lo que se refiere al aspecto temporal de esta relación, el índice se ha demostrado su validez como predictor de los precios futuros.
73. El uso de información de portales permite disponer de los resultados del índice de precios actualizado y que sería fácilmente adaptable a un proceso casi en tiempo real (frecuencia diaria). Esto ofrece una perspectiva de control del mercado inmediato inusitada en el ámbito de la vivienda.
74. La complejidad de la presente metodología plantea varios desafíos operativos para su puesta en práctica. El número de modelos, variables, fuentes y estratos involucrados implican un coste de mantenimiento operativo alto, resolubles con la automatización de los procesos de carga, tratamiento de la información y ajuste de modelos.

Las próximas líneas de investigación asociadas a la presente Tesis doctoral se centrarán en tres aspectos: 1) el uso de métodos hedónicos más precisos; 2) la solución de los sesgos de información ausente y zonal; y 3) la mejora de los métodos de tratamiento de series temporales. Más particularmente, atendiendo a los siguientes puntos:

a. Los recientes avances en el campo del aprendizaje profundo, y la limitada aplicación al campo de la valoración de la vivienda, permite plantear la aplicación de métodos de regresión basados en redes neuronales artificiales de aprendizaje profundo (*deep learning*). Más concretamente, se trabajará sobre las propuestas metodológicas del estado del arte en redes profundas

de grafos, y las de tipo espacio-temporal.

b. Analizar el impacto de incorporar o desarrollar nuevas fuentes de datos que complementen aquellos segmentos que tienen menor soporte de información, como las viviendas unifamiliares o los inmuebles singulares.

c. Replantear el proceso de reponderación, para que calibración incorpore los aspectos geográficos de los inmuebles con mayor detalle, de tal forma que se evite la aplicación de un modelo final de corrección de sesgos.

d. Incorporar procesos de identificación de anomalías en la estructura de las series de precios generadas, de forma que sea posible su corrección automática a través de mecanismos de imputación.

e. Plantear mejoras en el proceso de desagregación temporal y conciliación de series temporales con distinta frecuencia. Primeramente, con la inclusión de distintos aspectos zonales, temporales y funcionales en el proceso para asegurar la consistencia multinivel de las series; en segundo lugar, incorporar una aproximación multivariada para lograr resultados más robustos; y finalmente, incorporar funciones de densidad no paramétricas para estimar las funciones de verosimilitud en el algoritmo de selección del mejor método de descomposición.

En resumen, este trabajo contribuye al campo de la economía ofreciendo una serie de novedosos modelos de índices de precios hedónicos de la vivienda en alquiler, cuya innovación es aplicación conjunta de la estadística, econometría tradicional, el aprendizaje automático y las fuentes de datos de portales inmobiliarios.

Se han abordado eficazmente los retos en la incorporación del factor de la localización como determinantes de los precios; la solución de los problemas de coherencia y calidad de los datos inmobiliarios procedentes de portales; la especificación de una función de enlace entre precios de oferta y reales; y finalmente, el diseño temporal y poblacional de los resultados a través de un índice de precios altamente desagregado que guarda consistencia con los datos de la estadística oficial. Se espera que este trabajo sea una referencia para que permita sucesivas investigaciones para el avance del estado del arte en diferentes ámbitos como son la aplicación de técnicas no tradicionales en el modelado hedónico de la vivienda, o la incorporación de fuentes alternativas para complementar o desarrollar herramientas de información para la gestión pública.

Referencias bibliográficas

Abdi H. y Williams L.J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2 (4): 433-459.

Adamek J. (1994). Fusion: combining data from separate sources. Marketing Research 6 (3): 48.

Agostini C. y Palmucci G. (2017). Capitalización anticipada del metro de Santiago en el precio de las viviendas. El Trimestre Económico 75: 403-431.

Alba E. de (1988). Disaggregation and forecasting: A Bayesian analysis. Journal of Business & Economic Statistics 6 (2): 197-206.

Alexander Dietzel M., Braun N. y Schäfers W. (2014). Sentiment-based commercial real estate forecasting with Google search volume data. Journal of Property Investment & Finance 32 (6): 540-569.

Alfaro Navarro J.-L., Cano E., Alfaro Cortés E., Garcıéa N., Gámez M. y Larraz B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. Complexity 2020.

Alonso W. y others (1964). Location and land use. Toward a general theory of land rent. Location and land use. Toward a general theory of land rent.

Ambrose B.W., Eichholtz P. y Lindenthal T. (2013). House prices and fundamentals: 355 years of evidence. Journal of Money, Credit and Banking 45 (2–3): 477-491.

Andridge R.R. y Little R.J. (2010). A review of hot deck imputation for survey non-response. International statistical review 78 (1): 40-64.

Anenberg E. y Laufer S. (2017). A more timely house price index. Review of Economics and Statistics 99 (4): 722-734.

Anscombe F.J. (1960). Rejection of outliers. Technometrics 2 (2): 123-146.

Anselin L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. Agricultural economics 27 (3): 247-267.

Anselin L. y Griffith D.A. (1988). Do spatial effecfs really matter in regression analysis? Papers in Regional Science 65 (1): 11-34.

Anselin L. y Rey S.J. (2014). Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL. GeoDa Press LLC.

Antipov E.A. y Pokryshevskaya E.B. (2012). Mass appraisal of residential

apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications 39 (2): 1772-1778.

Anundsen A.K., Gerdrup K., Hansen F. y Kragh-Sørensen K. (2016). Bubbles and crises: The role of house prices and credit. Journal of Applied Econometrics 31 (7): 1291-1311.

Ardila D., Ahmed A. y Sornette D. (2021). Comparing ask and transaction prices in the Swiss housing market. Quantitative Finance and Economics 5 (1): 67-93.

Arlot S. y Celisse A. (2010). A survey of cross-validation procedures for model selection.

Arnott R. (2003). Com Tenancy rent control. Swedish economic policy review 10 (1): 89-134.

Arnott R. y Shevyakhova E. (2014). Tenancy rent control and credible commitment in maintenance. Regional Science and Urban Economics 47: 72-85.

Arribas-Bel D. y Fleischmann M. (2022). Spatial Signatures - Understanding (urban) spaces through form and function. Habitat International 128: 102641. https://doi.org/https://doi.org/10.1016/j.habitatint.2022.102641.

Arruñada B. (2022). Comentario a las nuevas regulaciones del alquiler. FEDEA, Colección Apuntes (2022–12).

Auer L. von y Wengenroth J. (2020). Consistent aggregation with superlative and other price indices. Journal of the Royal Statistical Society: Series A (Statistics in Society) 184 (2): 589-615.

Ayouba K., Breuillé M.-L., Grivault C. y Le Gallo J. (2020). Does Airbnb disrupt the private rental market? An empirical analysis for French cities. International Regional Science Review 43 (1–2): 76-104.

Ayuntamiento de Madrid (2022). Panel de indicadores de distritos y barrios de Madrid, estudio sociodemográfico. https://datos.madrid.es.

Azqueta Oyarzun D. (1994). Valoración económica de la calidad ambiental. Madrid, ES: McGraw-Hill.

Azzalini A. (2017). Statistical inference: Based on the likelihood. Routledge.

Bagnoli C. y Smith H. (1998). The theory of fuzz logic and its application to real estate valuation. Journal of Real Estate Research 16 (2): 169-200.

Bailey M.J., Muth R.F. y Nourse H.O. (1963). A regression method for real estate price index construction. Journal of the American Statistical Association 58 (304): 933-942.

Baldominos A., Blanco I., Moreno A.J., Iturrarte R., Bernárdez Ó. y Afonso C. (2018). Identifying real estate opportunities using machine learning. Applied sciences 8 (11): 2321.

Balk B.M. (1995). Axiomatic price index theory: a survey. International Statistical

Review/Revue Internationale de Statistique 69-93.

Balk B.M. (2008). Axioms, Tests, and Indices. Cambridge University Press, p. 53-139. https://doi.org/10.1017/CBO9780511720758.004.

Ball M. (1974). The determinants of relative house prices: A reply. Urban Studies 11 (2): 231-233.

Ball M.J. (1973). Recent empirical work on the determinants of relative house prices. Urban studies 10 (2): 213-233.

Banco Central Europeo (2022). Portal de estadístical del Banco Central Europeo. https://www.ecb.europa.eu/stats/html/index.en.html.

Banco de España (2019). Evolución reciente del mercado del alquiler de vivienda en España. Boletín Económico 3.

Banco de España (2020). Public intervention in the rental housing market: a review of international experience. Documentos ocasionales/Banco de España, 2002.

Banco de España (2021). Indicadores del Mercado de la vivienda.

Banco de España (2022a). Boletín estadístico del Banco de España. https://www.bde.es/webbde/es/estadis/infoest/bolest.html.

Banco de España (2022b). Nota metodológica de la Síntesis de Indicadores del Mercado Inmobiliario. Banco de España. https://www.bde.es/webbde/es/estadis/infoest/sindi.html.

Barbone L., Bodo G. y Visco I. (1981). Costi e profitti nell'industria in senso stretto: un'analisi su serie trimestrali, 1970-1980. Bollettino della Banca d'Italia 36: 465-510.

Barcellan R. (1994). Ecotrim: A program for temporal disaggregation of time series. En: Workshop on Quarterly National Accounts, Eurostat, Theme, Vol. 2. p. 79-95.

Barnett V. y Lewis T. (1984). Outliers in statistical data. osd.

Barron K., Kung E. y Proserpio D. (2021). The effect of home-sharing on house prices and rents: Evidence from Airbnb. Marketing Science 40 (1): 23-47.

Bartholomew K.A. y Ewing R. (2011). Hedonic Price Effects of Pedestrian- and Transit-Oriented Development. Journal of Planning Literature 26: 18-34.

Batty M. (2009). Accessibility: In search of a unified theory. Environment and Planning B: Planning and Design 36 (2): 191-194. https://doi.org/10.10620188/b3602ed.

Bax D., Zewotir T. y North D. (2021). Appraising residential property using hierarchical generalised additive models. Journal of Property Research 38 (3): 198-212.

Ben-Gal I. (2005). Outlier detection. En: Data mining and knowledge discovery handbook. Springer, p. 131-146.

Bennet T. (1920). The theory of measurement of changes in cost of living. Journal

of the Royal Statistical Society 83 (3): 455-462.

Berndt E.R. (1991). The measurement of quality change: constructing an hedonic price index for computers using multiple regression methods. The practice of econometrics: Classic and contemporary 102-149.

Berndt E.R. y Rappaport N.J. (2001). Price and quality of desktop and mobile personal computers: A quarter-century historical overview. American Economic Review 91 (2): 268-273.

Beullens K., Matsuo H., Loosveldt G. y Vandenplas C. (2014). Quality report for the European Social Survey, round 6. London: European Social Survey ERIC.

Biancotti C., Kirchner R., Mouriaux F., Rosolia A. y Veronese G. (2020). Covid-19 and official statistics: a wake-up call?

Bickel P.J. y Doksum K.A. (2015). Mathematical statistics: basic ideas and selected topics, volumes I-II package. Chapman; Hall/CRC.

Blanchard S.D. y Waddell P. (2017). Urbanaccess: generalized methodology for measuring regional accessibility with an integrated pedestrian and transit network. Transportation research record 2653 (1): 35-44.

Bloomberg L.N. (1947). Rent Control and the Housing Shortage: A Commentary on Roofs or Ceilings? by Friedman and Stigler. J. Land & Pub. Util. Econ. 23: 214.

Boeing G. (2020). Online rental housing market representation and the digital reproduction of urban inequality. Environment and Planning A: Economy and Space 52 (2): 449-468.

Boeing G., Higgs C., Liu S., Giles-Corti B., Sallis J.F., Cerin E., Lowe M., Adlakha D., Hinckson E., Moudon A.V. y others (2022). Using open data and open-source software to develop spatial indicators of urban design and transport features for achieving healthy and sustainable cities. The Lancet Global Health 10 (6): e907-e918.

Boeing G. y Waddell P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. Journal of Planning Education and Research 37 (4): 457-476.

Bondaruk B. (2019). Discrete Global Grid Systems: Operational Capability of the Current State of the Art.

Boot J. y Feibes W. (1967). On Glejser's derivation of monthly figures from yearly data. Brussels Economic Review 36: 589-596.

Borg I. y Groenen P.J. (2005). Modern multidimensional scaling: Theory and applications. Springer Science & Business Media.

Bourassa S.C., Hoesli M., Merlin L. y Renne J. (2021). Big data, accessibility and urban house prices. Urban Studies 58 (15): 3176-3195.

Bover O. y Velilla P. et al. (2001). Precios hedónicos de la vivienda sin

característicasticas: el caso de las promociones de viviendas nuevas. Estudios económicos.

Bowen W.M., Mikelbank B.A. y Prestegaard D.M. (2001). Theoretical and empirical considerations regarding space in hedonic housing price model applications. Growth and change 32 (4): 466-490.

Bowes D. y Ihlanfeldt K. (2001). Identifying the impacts of rail transit stations on residential property values. Journal of Urban Economics 50 (1): 1-25.

Box G.E., Jenkins G.M., Reinsel G.C. y Ljung G.M. (2016). Time series analysis: forecasting and control. Fifth edition. ed. Wiley series en probability y statistics. John Wiley & Sons.

Bram J. y Ludvigson S.C. (1998). Does consumer confidence forecast household expenditure? A sentiment index horse race. Economic Policy Review 4 (2).

Breiman L. (1996). Bagging predictors. Machine learning 24 (2): 123-140.

Breiman L. (2001). Random forests. Machine learning 45 (1): 5-32.

Breiman L. (2017). Classification and regression trees. Routledge.

Breiman L., Friedman J., Olshen R. y Stone C. (1984). Cart. Classification and regression trees.

Bricongne J.-C., Meunier B. y Pouget S. (2023). Web-scraping housing prices in real-time: The Covid-19 crisis in the UK. Journal of Housing Economics 59: 101906. https://doi.org/https://doi.org/10.1016/j.jhe.2022.101906.

Britannica (2014). Location theory. Encyclopedia Britannica. https://www.britannica.com/topic/location-theory.

Brunsdon C. y Comber A. (2021). Opening practice: supporting reproducibility and critical spatial data science. Journal of Geographical Systems 23 (4): 477-496.

Can A. (1992). Specification and estimation of hedonic housing price models. Regional science and urban economics 22 (3): 453-474.

Carbó Valverde S. y Rodriguez Fernandez F. (2010). The relationship between mortgage markets and house prices: does financial instability make the difference? Federal Reserve Bank of Atlanta CenFIS Working Paper 10-02.

Case B. y Quigley J.M. (1991). The dynamics of real estate prices. The Review of Economics and Statistics 50-58.

Case K.E. y others (1986). The market for single-family homes in the Boston area. New England Economic Review (May): 38-48.

Case K.E. y Quigley J.M. (2008). How housing booms unwind: income effects, wealth effects, and feedbacks through financial markets. European Journal of Housing Policy 8 (2): 161-180.

Case K.E. y Shiller R.J. (1987). Prices of single family homes since 1970: New indexes for four cities. National Bureau of Economic Research Cambridge, Mass., USA.

Case K.E. y Shiller R.J. (2003). Is there a bubble in the housing market? Brookings papers on economic activity 2003 (2): 299-362.

Casella G. y Berger R.L. (2021). Statistical inference. Cengage Learning.

Cassel C.-M., Sarndal C.-E. y Wretman J.H. (1977). Foundations of inference in survey sampling.

Cassel C., Särndal C. y Wretman J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63 (3): 615-620.

Cassel E. y Mendelsohn R. (1985). The choice of functional forms for hedonic price equations: comment. Journal of Urban Economics 18 (2): 135-142.

Causey B. y Trager M.L. (1981). Derivation of solution to the benchmarking problem: Trend revision. Unpublished research notes, US Census Bureau, Washington DC.

Čeh M., Kilibarda M., Lisec A. y Bajat B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS international journal of geo-information 7 (5): 168.

Cellmer R. (2023). Points of Interest and Housing Prices. Real Estate Management and Valuation 31 (1): 69-77. https://doi.org/doi:10.2478/remav-2023-0007.

Chapelle G. y Eymeoud J.B. (2022). Can big data increase our knowledge of local rental markets? A dataset on the rental sector in France. PloS one 17 (1): e0260405.

Chauvet M., Gabriel S. y Lutz C. (2013). Fear and loathing in the housing market: Evidence from search query data. Available at SSRN 2148769.

Chen B. y Andrews S. (2008). An empirical review of methods for temporal distribution and interpolation in the national accounts. Survey of current business 88 (5): 31.

Chen J. y Qin J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. Biometrika 80 (1): 107-116.

Chen J. y Sitter R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Statistica Sinica 385-406.

Chen R. y Tsay R. (1993). Nonlinear additive ARX models. Journal of the American Statistical Association 88 (423): 955-967.

Chen Z. y Dagum E. (1997). A recursive method for predicting variables with temporal and contemporaneous constraints. En: American Statistical Association, Proceedings of the Business and Economic Statistics Section. p. 229-233.

Cheshire P. y Sheppard S. (1995). On the Price of Land and the Value of Amenities. Economica 62: 247-267.

Chicco D., Warrens M.J. y Jurman G. (2021). The coefficient of determination R-

squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7: e623.

Choi K., Park H.J. y Uribe F.A. (2022). The impact of light rail transit station area development on residential property values in Calgary, Canada: Focus on land use diversity and activity opportunities. Case studies on transport policy 100924.

Cholette P.-A. (1984). Adjusting sub-annual series to yearly benchmarks. Statistics Canada, Methodology Branch, Time Series Research; Analysis ....

Cholette P.-A. (1988). Benchmarking systems of socio-economic time series. Statistics Canada, Methodology Branch, Time Series Research; Analysis ....

Cholette P. y Dagum E. (1994). Benchmarking time series with autocorrelated survey errors. International Statistical Review/Revue Internationale de Statistique 365-377.

Chow G. y Lin A. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. The review of Economics and Statistics 372-375.

Chung I.H. (2015). School choice, housing prices, and residential sorting: Empirical evidence from inter-and intra-district choice. Regional science and urban economics 52: 39-49.

Clapp J.M. (2004). A semiparametric method for estimating local house price indices. Real Estate Economics 32 (1): 127-160.

Clark S. y Lomax N. (2018). A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. Journal of big Data 5 (1): 1-21.

Cleveland W.S. y Devlin S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American statistical association 83 (403): 596-610.

Cleveland W.S., Devlin S.J. y Grosse E. (1988). Regression by local fitting: methods, properties, and computational algorithms. Journal of econometrics 37 (1): 87-114.

Consejo General del Poder Judicial - Nº 35 - junio 2013 (2012). Boletín Información Estadística. https://www.poderjudicial.com.

Corder G.W. y Foreman D. (2014). Nonparametric statistics: A step-by-step approach. John Wiley & Sons.

Court A.T. (1939). Hedonic price indexes with automotive examples.

Cousineau D. y Chartier S. (2010). Outliers detection and treatment: a review. International Journal of Psychological Research 3 (1): 58-67.

Curry B., Morgan P. y Silver M. (2002). Neural networks and non-linear statistical methods: an application to the modelling of price–quality relationships.

Computers & Operations Research 29 (8): 951-969.

d'Acci L. (2019). Quality of urban area, distance from city centre, and housing value. Case study on real estate values in Turin. Cities 91: 71-92.

D'Orazio M., Di Zio M. y Scanu M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.

Daalmans J. y Di Fonzo T. (2014). Denton PFD and GRP benchmarking are friends. An empirical evaluation on Dutch Supply and Use Tables. En: Unpublished paper presented at the 22nd International Input-Output conference.

Daalmans J., Di Fonzo T., Mushkudiani N. y Bikker R. (2018). Growth Rates Preservation (GRP) temporal benchmarking: Drawbacks and alternative solutions. Survey Methodology 44 (1): 43-61.

Dagum E. y Cholette P. (2006b). Reconciliation and Balancing Systems of Time Series. Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series 263-284.

Dagum E. y Cholette P. (2006a). Benchmarking, temporal distribution, and reconciliation methods for time series. Springer Science & Business Media.

Davies J., Elliot D., Aston J. y Sayal H. (2015). A comparison of new and established benchmarking methods. Working Paper). Cambridge, UK: CMIH. Retrieved from http://www. ccimi. maths . . . .

De Wit E.R., Englund P. y Francke M.K. (2013). Price and transaction volume in the Dutch housing market. Regional Science and Urban Economics 43 (2): 220-241.

DeGroot M.H. y Schervish M.J. (2012). Probability and statistics. Pearson Education.

DeJong D.N., Nankervis J.C., Savin N.E. y Whiteman C.H. (1992). Integration versus trend-stationarity in time series. Econometrica 60 (2): 423-433.

Del Cacho C. (2010). A comparison of data mining methods for mass real estate appraisal.

Delignette-Muller M.L., Dutang C. y others (2015). fitdistrplus: An R package for fitting distributions. Journal of statistical software 64 (4): 1-34.

Demchenko Y., De Laat C. y Membrey P. (2014). Defining architecture components of the Big Data Ecosystem. En: 2014 International conference on collaboration technologies and systems (CTS). IEEE, p. 104-112.

Deming W.E. y Stephan F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11 (4): 427-444.

Dempster A.P., Laird N.M. y Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39 (1): 1-22.

Deng Y., Ross S.L. y Wachter S.M. (2003). Racial differences in homeownership:

the effect of residential location. Regional Science and Urban Economics 33 (5): 517-556.

Denton F.T. (1971). Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. Journal of the American Statistical Association 66 (333): 99-102.

Des Rosiers F. y Thériault M. (2006). Mass appraisal, hedonic price modelling and urban externalities: Understanding property value shaping processes.

Desormeaux D. y Piguillem F. (2003). Precios hedónicos e índices de precios de viviendas. Documento de trabajo (12).

Devaud D. y Tillé Y. (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. Test 28 (4): 1033-1065.

Deville J.-C. (2000). Generalized calibration and application to weighting for non-response. En: COMPSTAT. Springer, p. 65-76.

Deville J.-C. y Särndal C.-E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association 87 (418): 376-382.

Di Fonzo T. (2002). Temporal Disaggregation of Economic Time Series: Towards a Dynamic Extension. European Commission (Eurostat) Working Papers and Studies, Theme 1, General Statistics (pp. 41).

Di Fonzo T. y Filosa R. (1987). Methods of estimation of quarterly national account series: a comparison. Journée franco-italienne de Compatibilité Nationale.

Di Fonzo T. y Marini M. (2005). Benchmarking a system of time series: Denton's movement preservation principle vs. a data based procedure. En: Proceedings of the Workshop in Frontiers in Benchmarking Techniques and their Application in Official Statistics, Luxembourg, Eurostat (to appear).

Di Fonzo T. y Marini M. (2011). Simultaneous and two-step reconciliation of systems of time series: methodological and practical issues. Journal of the Royal Statistical Society: Series C (Applied Statistics) 60 (2): 143-164.

Diamond R., McQuade T. y Qian F. (2019). The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco. American Economic Review 109 (9): 3365-94.

Diaz A. y Jerez B. (2013). House prices, sales, and time on the market: A search-theoretic framework. International Economic Review 54 (3): 837-872.

Díaz J.C. (1997). La teoría de los índices de precios. Cuadernos de estudios empresariales (7): 71-88.

Diewert E. y Shimizu C. (2021). Residential Property Price Indexes: Spatial Coordinates Versus Neighborhood Dummy Variables. Review of Income and Wealth.

Diewert W.E. (1976). Exact and superlative index numbers. Journal of econometrics 4 (2): 115-145.

Diewert W.E. (2003). Hedonic regressions: A review of some unresolved issues. En: 7th meeting of the Ottawa Group, Paris, May, Vol. 29.

Diewert W.E. (2005). Index number theory using differences rather than ratios. American Journal of Economics and Sociology 64 (1): 311-360.

Diewert W.E. (2009). The Paris OECD-IMF workshop on real estate price indexes: conclusions and future directions. Price and Productivity Measurement 1: 87-116.

Diewert W.E., Nishimura K.G., Shimizu C., Watanabe T. y others (2020). Property Price Index. Advances in Japanese Business and Economics.

Diewert W.E. y others (2007). Index numbers. Department of Economics, University of British Columbia.

Direccion General del Catastro (2020). Estadística del Catastro Inmobiliario Urbano. http://www.catastro.minhap.gob.es/esp/estadisticas.asp.

Direccion General del Catastro (2022). Registro Central del Catastro (Sede Electrónica). https://www.sedecatastro.gob.es.

Dixon W.J. (1950). Analysis of extreme values. The Annals of Mathematical Statistics 21 (4): 488-506.

Dolls M., Fuest C., Neumeier F. y Stöhlker D. (2021). Ein Jahr Mietendeckel: Wie hat sich der Berliner Immobilienmarkt entwickelt? ifo Schnelldienst 74 (3): 26-29.

Dubin R.A. (1998). Spatial autocorrelation: A primer. Journal of Housing Economics 7 (4): 304-327.

Echaves García A. y Martínez del Olmo A. (2021). Emancipación residencial y acceso de los jóvenes al alquiler en España: Un problema agravado y su diversidad territorial. Ciudad y Territorio Estudios Territoriales 53 (M): 27-42. https://doi.org/10.37230/CyTET.2021.M21.02.

Edgeworth F.Y. (1888). Some new methods of measuring variation in general prices. Journal of the Royal Statistical Society 51 (2): 346-368.

Edgeworth F.Y. (1925). The Plurality of Index-Numbers. The Economic Journal 35 (139): 379-388.

Elfayoumi K., Salas J. y Tudyka A. (2021). Affordable Rental Housing: Making It Part of Europe’s Recovery. Departmental Papers 2021 (13).

Empirica (2021). Housing market data available for Q1 2021. https://www.empirica-regio.de/en/news/.

Enders W. (2014). Applied econometric time series. 4th ed. ed. Wiley Series en Probability y Statistics. J. Wiley, Hoboken, NJ.

European Statistical System (2023). https://ec.europa.eu/eurostat/web/european-statistical-system.

Eurostat (2013). Handbook on quarterly national accounts.

Eurostat (2014). Handbook on Residential Property Prices (RPPIs). En: Statistical Office of the European Communities and International Labour Office. International Monetary Fund Washington, DC.

Eurostat (2015). ESS guidelines on temporal disaggregation : benchmarking and reconciliation.

Eurostat (2017a). Technical manual on Owner-Occupied Housing and House Price Indices. 138. http://ec.europa.eu/eurostat/documents/7590317/0/Technical-Manual-OOH-HPI-2017/.

Eurostat (2017b). HICP Recommendation on Obtaining Scanner Data.

Eurostat (2021). Housing in Europe. https://ec.europa.eu/eurostat/cache/digpub/housing/index.html?lang=en.

Eurostat (2022). Housing price statistics - house price index sales. Eurostat. https://ec.europa.eu/eurostat/cache/metadata/en/prc_hpi_inx_esms.htm.

Fan G.-Z., Ong S.E. y Koh H.C. (2006). Determinants of house price: A decision tree approach. Urban Studies 43 (12): 2301-2315.

Fenwick D. (2013). Uses of residential property price indices.

Fernandez R. (1981). A methodological note on the estimation of time series. The Review of Economics and Statistics 63 (3): 471-476.

Finect (2021). Deducción por alquiler en la renta. https://www.finect.com/usuario/Josetrecet/articulos/deduccion-alquiler-renta.

Fisher I. (1922a). The making of index numbers: a study of their varieties, tests, and reliability. N.º 1. Houghton Mifflin.

Fisher R.A. (1922b). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A 222: 309-368.

Fix E. y Hodges J.L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique 57 (3): 238-247.

Fletcher M., Gallimore P. y Mangan J. (2000). Heteroscedasticity in hedonic house price models. Journal of Property Research 17 (2): 93-108.

Follain J.R. y Malpezzi S. (1980). Dissecting housing value and rent: Estimates of hedonic indexes for thirty-nine large SMSAs. Vol. 249. Urban Institute Press.

Folsom R.E. y Singh A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. En: Proceedings of the American Statistical Association, Survey Research Methods Section, Vol. 598603.

Fonzo T.D. y Marini M. (2013). Benchmarking and movement preservation: Evidences from real-life and simulated series. En: Advances in theoretical and applied statistics. Springer, p. 499-509.

Fotocasa (2017). Metodología de cálculo del índice de precios de la vivienda en alquiler. https://prensa.fotocasa.es/wp-content/uploads/2017/12/Metodologia-indice-alquiler.pdf.

Fotocasa (2021). Número de anuncios publicados en fotocasa. https://www.fotocasa.es/.

Fox J. (2000). Multiple and generalized nonparametric regression. Vol. 7. Sage.

Franco S.F. y Santos C.D. (2021). The impact of Airbnb on residential property values and rents: Evidence from Portugal. Regional Science and Urban Economics 88: 103667.

Frank L.D., Sallis J.F., Saelens B.E., Leary L., Cain K., Conway T.L. y Hess P.M. (2010). The development of a walkability index: application to the Neighborhood Quality of Life Study. British journal of sports medicine 44 (13): 924-933.

Freeman M. (1979). The hedonic price approach to measuring demand for neighborhood characteristics. 191-217.

Freund Y., Schapire R. y Abe N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14 (771–780): 1612.

Friedman J. y Weinberg D.H. (1981). The demand for rental housing: Evidence from the housing allowance demand experiment. Journal of Urban Economics 9 (3): 311-331.

Fuller W.A. (2011). Sampling statistics. John Wiley & Sons.

Füss R. y Koller J.A. (2016). The role of spatial and temporal structure for residential rent predictions. International Journal of Forecasting 32 (4): 1352-1368.

Galesi A., Mata N., Rey D., Schmitz S. y Schuffels J. (2020). Regional housing market conditions in spain. Available at SSRN 3724178.

Garcia-López M.-À., Jofre-Monseny J., Martínez-Mazza R. y Segú M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. Journal of Urban Economics 103278.

Gasparini L. y Sosa Escudero W. (1999). Bienestar y distribución del ingreso en Argentina, 1980-1998. Económica 45.

Ge J., Runeson G. y Lam K. (2003). Forecasting Hong Kong housing prices: An artificial neural network approach. En: International conference on methodologies in housing research, Stockholm, Sweden.

Giffen B. van, Herhausen D. y Fahse T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research 144: 93-106.

Giuliano G. y Small K. (1991). Subcenters in the Los Angeles region.

Goh Y.M., Costello G. y Schwann G. (2012). Accuracy and robustness of house price

index methods. Housing Studies 27 (5): 643-666.

Gómez-Rubio V. (2020). Bayesian inference with INLA. CRC Press.

Goodman A.C. y Thibodeau T.G. (1995). Age-related heteroskedasticity in hedonic house price. Journal of Housing Research 6: 25-42.

Goodman A.C. y Thibodeau T.G. (1998). Housing Market Segmentation. Journal of Housing Economics 7 (2): 121-143. https://doi.org/10.1006/jhec.1998.0229.

Graczyk M., Lasota T. y Trawiński B. (2009). Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. 800-812.

Graczyk M., Lasota T., Trawiński B. y Trawiński K. (2010). Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. 340-350.

Graf B. (2020). Consumer Price Index Manual, 2020: Concepts and Methods. En: Consumer Price Index Manual, 2020. International Monetary Fund.

Griliches Z. (1961). Hedonic price indexes for automobiles: An econometric of quality change. En: The price statistics of the federal goverment. NBER, p. 173-196.

Griliches Z. (1990). Hedonic price indexes and the measurement of capital and productivity: some historical reflections. 185-202.

Grömping U. (2009). Variable importance assessment in regression: linear regression versus random forest. The American Statistician 63 (4): 308-319.

Guerrero V.M. y Martínez J. (1995). A recursive ARIMA-based procedure for disaggregating a time series variable using concurrent data. Test 4 (2): 359-376.

Guilkey D., Miles M. y Cole R. (1989). The motivation for institutional real estate sales and implications for asset class returns. Real Estate Economics 17 (1): 70-86.

Guyet T., Spillemaecker L., Malinowski S. y Graux A.-I. (2022). Temporal Disaggregation of the Cumulative Grass Growth. 383-394.

Han L. y Strange W. (2016). What is the role of the asking price for a house? Journal of Urban Economics 93: 115-130.

Han L. y Strange W.C. (2014). Bidding wars for houses. Real Estate Economics 42 (1): 1-32.

Handy S. (2020). Is accessibility an idea whose time has finally come? Transportation Research Part D: Transport and Environment 83: 102319.

Handy S.L. y Niemeier D.A. (1997). Measuring accessibility: an exploration of issues and alternatives. Environment and planning A 29 (7): 1175-1194.

Hanink D., Cromley R. y Ebenstein A. (2012). Spatial variation in the determinants of house prices and apartment rents in China. The Journal of Real Estate Finance and Economics 45 (2): 347-363.

Hansen L.K. y Salamon P. (1990). Neural network ensembles. IEEE transactions

on pattern analysis and machine intelligence 12 (10): 993-1001.

Hansen W.G. (1959). How Accessibility Shapes Land Use. Journal of the American Planning Association 25 (2): 73-76. https://doi.org/10.1080/01944365908978307.

Hanushek E.A. y Quigley J.M. (1979). The dynamics of the housing market: A stock adjustment model of housing consumption. Journal of Urban Economics 6 (1): 90-111.

Härdle W. y Linton O. (1994). Applied nonparametric methods. Handbook of econometrics 4: 2295-2339.

Hashem S. (1997). Optimal linear combinations of neural networks. Neural networks 10 (4): 599-614.

Hastie T. y Tibshirani R. (2017). Generalized additive models. Routledge.

Hastie T., Tibshirani R. y Friedman J. (2017). The elements of statistical learning: data mining, inference, and prediction. Second Edition. Springer Science & Business Media.

Hawkins D.M. (1980). Identification of outliers. Vol. 11. Springer.

He Z., Xu X. y Deng S. (2003). Discovering cluster-based local outliers. Pattern Recognition Letters 24 (9–10): 1641-1650.

Heikkila E., Gordon P., Kim J.I., Peiser R.B., Richardson H.W. y Dale-Johnson D. (1989). What Happened to the CBD-Distance Gradient?: Land Values in a Policentric City. Environment and Planning A 21 (2): 221-232.

Helbich M., Brunauer W., Vaz E. y Nijkamp P. (2014). Spatial heterogeneity in hedonic house price models: The case of Austria. Urban Studies 51 (2): 390-411.

Henderson J.V. (1985). The impact of zoning policies which regulate housing quality. Journal of Urban Economics 18 (3): 302-312.

Heyman A., Law S. y Berghauser Pont M. (2018). How is Location Measured in Housing Valuation? A Systematic Review of Accessibility Specifications in Hedonic Price Models. Urban Science 3 (1): 3. https://doi.org/10.3390/urbansci3010003.

Heyman A.V. y Sommervoll D.E. (2019). House prices and relative location. Cities 95: 102373.

Hill R.J. (2006). When does chaining reduce the Paasche–Laspeyres spread? An application to scanner data. Review of Income and Wealth 52 (2): 309-325.

Hill R.J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. Journal of economic surveys 27 (5): 879-914.

Hill R.J. y Scholz M. (2018). Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines. Review of Income and Wealth 64: 737-756. https://doi.org/10.1111/roiw.12303.

Hill R., Scholz M., Shimizu C. y Steurer M. (2018). An evaluation of the methods used by European countries to compute their official house price indices. Economie et Statistique 500 (1): 221-238.

Hillier B. y Hanson J. (1989). The social logic of space. Cambridge university press.

Hillmer S.C. y Trabelsi A. (1987). Benchmarking of economic time series. Journal of the American Statistical Association 82 (400): 1064-1071.

Hjort A., Pensar J., Scheel I. y Sommervoll D.E. (2022). House price prediction with gradient boosted trees under different loss functions. Journal of Property Research 1-27.

Ho W., Tang B.-S. y Wong S. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research 38 (1): 48-70.

Hofsten E.A.G. von y others (1952). Price indexes and quality changes. Bokforlage Forum.

Hong J., Choi H. y Kim W. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management 24 (3): 140-152.

Hood C.C.H. (2005). An empirical comparison of methods for benchmarking seasonally adjusted series to annual totals. Proceedings of the American Statistical Association, Business and Economic Statistics Section.[CD-ROM] http://www. census. gov/ts/papers/chood_asa2005. pdf.

Horn K. y Merante M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. Journal of Housing Economics 38: 14-24.

Hornik K., Stinchcombe M. y White H. (1989). Multilayer feedforward networks are universal approximators. Neural networks 2 (5): 359-366.

Horowitz J.L. y Lee S. (2002). Semiparametric methods in applied econometrics: Do the models fit the data? Statistical Modelling 2 (1): 3-22.

Hort M., Chen Z., Zhang J.M., Sarro F. y Harman M. (2022). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. https://arxiv.org/abs/2207.07068.

Hu L., He S. y Su S. (2022). A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. Computers, Environment and Urban Systems 94: 101775.

Hu Y. y Han Y. (2019). Identification of urban functional areas based on POI data: A case study of the Guangzhou economic and technological development zone. Sustainability 11 (5): 1385.

Hubert M. y Vandervieren E. (2008). An adjusted boxplot for skewed distributions. Computational statistics & data analysis 52 (12): 5186-5201.

Hwang M. y Quigley J.M. (2010). Housing price dynamics in time and space: predictability, liquidity and investor returns. The Journal of Real Estate Finance

and Economics 41: 3-23.

Hyndman R.J. y Athanasopoulos G. (2018). Forecasting: principles and practice. OTexts.

Idealista (2019). Metodología de cálculo del índice de precios de la vivienda en alquiler. https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/.

Idealista (2020). Informe anual idealista/data 2020. http://www.idealista.com/data.

Idealista (2021). Número de anuncios publicados en idealista. https://www.idealista.com/.

Idealista (2022). Datos del informe de precios del mercado inmobiliario de Idealista - Diciembre 2022. http://www.idealista.com/data.

INE (2006b). Indice de Precios de la vivienda - Base 2007. https://www.ine.es/daco/daco42/ipv/metodologia.pdf.

INE (2006a). Metodología de la Encuesta de presupuestos familiares. Base 2006.

INE (2009). Metodología Indice del Precio de la Vivienda, base 2007.

INE (2011). Metodología del Censo de viviendas y Edificios 2011.

INE (2016b). Clasificación de bienes y servicios ECOICOP. https://www.ine.es/dynt3/inebase/index.htm?padre=3929&capsel=3929.

INE (2016a). Metodología Indice del Precio de la Vivienda, base 2015.

INE (2017). Metodología del Índice de Precios de Consumo. Base 2016.

INE (2019). Metodología: Encuesta de Condiciones de Vida. Instituto Nacional de Estadística: Madrid, Spain.

INE (2021a). Encuesta contínua de hogares. Resultados nacionales. https://www.ine.es/dynt3/inebase/index.htm?type=pcaxis&path=/t20/p274/serie/prov/p01&file=pcaxis&L=0&dh=0&capsel=0.

INE (2021b). Índice de Precios de la Vivienda en Alquiler.

INE (2022d). Estudio piloto de movilidad a partir del posicionamiento de teléfonos móviles.

INE (2022a). Datos de contabilidad Nacional - Diciembre 2022.

INE (2022c). Datos Índice de Precios de Consumo (IPC) - Diciembre 2022.

INE (2022b). Datos del Índice de Precios de la Vivienda - Diciembre 2022. https://www.ine.es/prensa/ipv_tabla1.htm.

INE (2023d). Estadística continua de población. Resultados. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177095&menu=resultados&idp=1254735572981.

INE (2023a). Encuesta de ocupación en hotelera. EOAT. Julio 2023. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176962&menu=ultiDatos&idp=1254735576863.

INE (2023e). Viviendas turísticas en España. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176962&menu=ultiDatos&

idp=1254735576863.

INE (2023b). Estadísticas de hipotecas. Estadísticas Nacionales.

INE (2023c). Encuesta de Población activa.

Internacional F.M. (2003). World Economic Outlook - April 2003 - Growth and Institutions. En: IMF eLibrary. International Monetary Fund.

Internacional F.M. (2022). World Economic Outlook - Julio 2022 - Gloomy and More Uncertain. En: IMF eLibrary. International Monetary Fund.

Jarque C.M. y Bera A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics letters 6 (3): 255-259.

Jarrell M.G. (1992). A comparison of two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, for identifying multivariate outliers.

Judd C.M., McClelland G.H. y Ryan C.S. (2011). Data analysis: A model comparison approach. Routledge.

Kadane J. (1978). Some statistical problems in merging data files. 1978 Compendium of Tax Research 17: 159-171.

Kain J.F. y Quigley J.M. (1970). Measuring the value of housing quality. Journal of the American statistical association 65 (330): 532-548.

Kaiser H.F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika 23 (3): 187-200.

Katranji M., Thuillier E., Kraiem S., Moalic L. y Selem F.H. (2016). Mobility data disaggregation: A transfer learning approach. 1672-1677.

Kauko T., Hooimeijer P. y Hakfoort J. (2002). Capturing housing market segmentation: An alternative approach based on neural network modelling. Housing Studies 17 (6): 875-894.

Kestens Y., Thériault M. y Des Rosiers F. (2006). Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles. Journal of Geographical Systems 8 (1): 61-96.

Kholodilin K. (2020). Long-Term, Multicountry Perspective on Rental Market Regulations. Housing Policy Debate 30 (6): 994-1015. https://doi.org/10.1080/10511482.2020.1789889.

Kiel K.A. y Zabel J.E. (2008). Location, location, location: The 3L Approach to house price determination. Journal of Housing Economics 17 (2): 175-190.

Knaap G.J. y Song Y. (2003). New urbanism and housing values: a disaggregate assessment. Vol. 54. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.7545%7B/&%7Drep=rep1%7B/&%7Dtype=pdf.

Kokot S. y Bas M. (2015). The comparative analysis of asking and traded price indices in different floor area subsegments of the residential property market. Real Estate Management and Valuation 23 (3): 14-25.

Kolbe J., Schulz R., Wersing M. y Werwatz A. (2021). Real estate listings and their

usefulness for hedonic regressions. Empirical economics 61 (6): 3239-3269.

Kontrimas V. y Verikas A. (2011). The mass appraisal of the real estate by computational intelligence. Applied Soft Computing 11 (1): 443-448.

Konüs A.A. (1924). The problem of the true index of the cost of living. Ekonomicheskaya Zhizn.

Koster H., Van Ommeren J. y Volkhausen N. (2018). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles.

Kott P. y Liao D. (2017). Calibration weighting for nonresponse that is not missing at random: Allowing more calibration than response-model variables. Journal of Survey Statistics and Methodology 5 (2): 159-174.

Krogh A. (2008). What are artificial neural networks? Nature biotechnology 26 (2): 195-197.

Krogh A. y Vedelsby J. (1994). Neural network ensembles, cross validation, and active learning. Advances in neural information processing systems 7.

Kuhn M., Johnson K. y others (2018). Applied predictive modeling, 2nd edition. Vol. 26. Springer.

Kullback L. (2012). On Information and Sufficiency. 22 (1): 79-86.

La Paz P.T. de y others (2021). Predicting housing prices. A long term housing price path for Spanish regions. Latin American Real Estate Society (LARES).

Lacerda N. (2018). Mercado imobiliário de bens patrimoniais: um modelo interpretativo a partir do centro histórico do Recife (Brasil). EURE (Santiago) 44 (132): 89-108.

Lang M., Binder M., Richter J., Schratz P., Pfisterer F., Coors S., Au Q., Casalicchio G., Kotthoff L. y Bischl B. (2019). mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software. https://doi.org/10.21105/joss.01903.

Larraz B. y Poblacion J. (2013). An online real estate valuation model for control risk taking: A spatial approach. Investment Analysts Journal 2013 (78): 83-96.

LeCun Y., Bengio Y. y Hinton G. (2015). Deep learning. nature 521 (7553): 436-444.

LeSage J. y Pace R.K. (2009). Introduction to spatial econometrics. Chapman; Hall/CRC.

Leucescu A. y Agafitei M. (2013). Statistical matching: a model based approach for data integration. Eurostat methodologies and Working papers. https://doi.org/10.2785/44822.

Levinson D. y Krizek K. (2005). Access to destinations. Elsevier Publishers.

Li H., Wei Y.D., Wu Y. y Tian G. (2019). Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure. Cities 91: 165-179.

Li K.-C. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. The Annals of Statistics 230-240.

Li Z. y Wood S.N. (2020). Faster model matrix crossproducts for large generalized linear models with discretized covariates. Statistics and Computing 30 (1): 19-25.

Liaw A. y Wiener M. (2002). Classification and regression by randomForest. R news 2 (3): 18-22.

Lieske S.N., Nouwelant R. van den, Han J.H. y Pettit C. (2021). A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices. Urban Studies 58 (1): 182-202.

Lin Y. y Zhang H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics 34 (5): 2272-2297.

Lisman J.H.C. y Sandee J. (1964). Derivation of quarterly figures from annual data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 13 (2): 87-90.

Litterman R. (1983). A random walk, Markov model for the distribution of time series. Journal of Business & Economic Statistics 1 (2): 169-173.

Liu J.-G., Zhang X.-L. y Wu W.-P. (2006). Application of fuzzy neural network for real estate prediction. En: International Symposium on Neural Networks. Springer, p. 1187-1191.

Liu S., Higgs C., Arundel J., Boeing G., Cerdera N., Moctezuma D., Cerin E., Adlakha D., Lowe M. y Giles-Corti B. (2022). A generalized framework for measuring pedestrian accessibility around the world using open data. Geographical Analysis 54 (3): 559-582.

Lloyd S. (1982). Least squares quantization in PCM. IEEE transactions on information theory 28 (2): 129-137.

Loberto M., Luciani A., Pangallo M. y others (2018). The potential of big housing data: an application to the Italian real-estate market. Banca d'Italia, Eurosistema.

Löchl M. (2010). Application of spatial analysis methods for understanding geographic variation of prices, demand and market success. ETH Zurich.

Lohr S.L. (2019). Sampling: design and analysis. Chapman; Hall/CRC.

López J. (2007). Los índices de precio de la vivienda. Problemática. Revista Índice, Revista de Estadística y Sociedad 14-17.

Lowe J. (1824). The present state of England in regard to agriculture, trade and finance: with a comparison of the prospects of England and France. E. Bliss; E. White.

Lundberg S.M., Erion G.G. y Lee S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lundberg S.M. y Lee S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Malpezzi S. y others (2003). Hedonic pricing models: a selective and applied review. Housing economics and public policy 1: 67-89.

Martínez S. y Rueda Rueda M. (2002). Estimadores de calibración: una nueva metodología para el uso de la información auxiliar. Metodología de encuestas 4 (2): 161-174.

McCluskey W. y Anand S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties. Journal of Property Investment & Finance.

McCulloch W. y Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5: 115-133.

McLaughlin R. y Young C. (2018). Data democratization and spatial heterogeneity in the housing market. A Shared Future: Fostering Communities of Inclusion in an Era of Inequality. Cambridge, MA: Harvard Joint Center for Housing Studies 126-139.

Meinshausen N. (2006). Quantile regression forests. Journal of Machine Learning Research 7 (Jun): 983-999.

Meng X.-L. y Rubin D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. En: Biometrika, Vol. 80. http://biomet.oxfordjournals.org/.

Meyer H., Reudenbach C., Wöllauer S. y Nauss T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecological Modelling 411: 108815.

Miller J.N. (1993). Tutorial review—Outliers in experimental data and their treatment. Analyst 118 (5): 455-461.

Mills E.S. (1972). Studies in the Structure of the Urban Economy.

Ministerio de la Gobernación (1944). Orden de 29 de febrero de 1944, por la que se determinan las condiciones higiénicas mínimas que han de reunir las viviendas. N.º BOE-A-1944-2079. https://www.boe.es/buscar/doc.php?id=BOE-A-1944-2079.

Miralles J.M.P. (1997). La problemática de la trimestralización de series anuales. Universitat de Valencia (Spain).

MITMA (2020). Sistema Estatal de Índices de Referencia del Precio del Alquiler de Vivienda. http://www.fomento.gob.es/be2/?nivel=2&orden=34000000.

MITMA (2022a). Estadísticas del sector de la construcción, Vivienda y actuaciones urbanas. http://www.fomento.gob.es/be2/?nivel=2&orden=34000000 .

MITMA (2022b). Observatorio de vivienda y suelo. Ministerio de Transportes, Movilidad y Agenda Urbana. https://apps.mitma.gob.es/CVP/.

Moauro F. y Savio G. (2005). Temporal disaggregation using multivariate structural time series models. The Econometrics Journal 8 (2): 214-234.

https://doi.org/10.1111/j.1368-423x.2005.00161.x.

Monahan J. (2011). Numerical methods of statistics. Cambridge University Press.

Monràs J. y Montalvo J.G. (2022). The effect of second generation rent controls: New evidence from Catalonia. Department of Economics; Business, Universitat Pompeu Fabra.

Montalvo J.G. (2011). De la quimera inmobiliaria al colapso financiero. Antoni Bosch Editor.

Montanari G.E. y Ranalli M.G. (2005). Nonparametric model calibration estimation in survey sampling. Journal of the American Statistical Association 100 (472): 1429-1442.

Montero J.M., Fernández-Avilés G. y Mateu J. (2015). Spatial and spatio-temporal geostatistical modeling and kriging. Wiley series en probability y statistics. Wiley, Chichester, West Sussex.

Montgomery J. (1929). Is There a Theoretically Correct Price Index of a Group of Commodities? L'Universale Tipogr. poliglotta.

Moralı O. y Yılmaz N. (2020). An analysis of spatial dependence in real estate prices. The Journal of Real Estate Finance and Economics 1-23.

Münger B. (2021). Generalized Additive Model Implementation for Germany Real Estate Market-Model, API, UI Development.

Muth R.F. (1969). Cities and housing, the spatial pattern of urban residential land use.

Nguyen y Cripps (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. Journal of real estate research 22 (3): 313-336.

Obe R. y Hsu L. (2011). PostGIS in action. GEOInformatics 14 (8): 30.

OCDE (2018). Housing prices. https://data.oecd.org/price/housing-prices.htm.

Ohnishi T., Mizuno T., Shimizu C. y Watanabe T. (2011). On the evolution of the house price distribution.

Olsen E.O. (1988). What do economists know about the effect of rent control on housing maintenance? The journal of real estate finance and economics 1 (3): 295-307.

OpenStreetMap (2017). Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

Opitz D.W. y Shavlik J.W. (1996). Actively searching for an effective neural network ensemble. Connection Science 8 (3–4): 337-354.

Orford S. (2017). Valuing the built environment: GIS and house price analysis. Routledge.

Orr J.M., Sackett P.R. y Dubois C.L. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration.

Personnel Psychology 44 (3): 473-486.

Osborne J. y Overbay A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation 9 (1): 6.

Osborne J.W., Christianson W.R. y others (2001). Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field.

Ottensmann J.R., Payton S. y Man J. (2008). Urban location and housing prices within a hedonic model. Journal of Regional Analysis and Policy 38 (1).

Owusu-Ansah A. (2011). A review of hedonic pricing models in housing research. Journal of International Real Estate and Construction Studies 1 (1): 19.

Özsoy O. y Şahin H. (2009). Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model. International Journal of Housing Markets and Analysis.

Paass G. (1986). Statistical match: evaluation of existing procedures and improvements by using additional information. Microanalytic Simulation Models to Support Social and Financial Policy 401-420.

Pace K. (1998). Appraisal using generalized additive models. Journal of Real Estate Research 15 (1): 77-99.

Pace R.K. (1995). Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models. The Journal of Real Estate Finance and Economics 11 (3): 195-217.

Páez A., Long F. y Farber S. (2008). Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. Urban Studies 45 (8): 1565-1581.

Pagano T.P., Loureiro R.B., Lisboa F.V.N., Cruz G.O.R., Peixoto R.M., Sousa Guimarães G.A. de, Santos L.L. dos, Araujo M.M., Cruz M., Oliveira E.L.S. de, Winkler I. y Nascimento E.G.S. (2022). Bias and unfairness in machine learning models: a systematic literature review. https://arxiv.org/abs/2202.08176.

Paige C.C. (1979). Fast numerically stable computations for generalized linear least squares problems. SIAM Journal on Numerical Analysis 16 (1): 165-171.

Palmquist R.B. (1989). Land as a differentiated factor of production: A hedonic model and its implications for welfare measurement. Land economics 65 (1): 23-28.

Pangallo M. y Loberto M. (2018). Home is where the ad is: online interest proxies housing demand. EPJ Data science 7 (1): 47.

Pareja-Eastaway M. y Sánchez-Martínez T. (2017). Social housing in Spain: what role does the private rented market play? Journal of Housing and the Built Environment 32 (2): 377-395. https://doi.org/10.1007/s10901-016-9513-6.

Pearson K. (1901). LIII. On lines and planes of closest fit to systems of points in

space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11): 559-572. https://doi.org/10.1080/14786440109462720.

Pérez-Rave J.I., Correa-Morales J.C. y González-Echavarrıéa F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research 36 (1): 59-96.

Pérez-Villalta R.A. (2002). ?‘ Qué es un modelo de superpoblación? Metodología de Encuestas 4 (1): 79-86.

Piazzesi M., Schneider M. y Stroebel J. (2015). Segmented housing search. National Bureau of Economic Research.

Pollakowski H.O. (1995). Data sources for measuring house price changes. Journal of Housing Research 377-387.

Pow N., Janulewicz E. y Liu L. (2014). Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal. Course project, COMP-598, Fall/2014, McGill University.

Prinzie A. y Van den Poel D. (2008). Random forests for multiclass classification: Random multinomial logit. Expert systems with Applications 34 (3): 1721-1732.

Quigley J.M. (1995). A simple hybrid model for estimating real estate price indexes. Journal of Housing Economics 4 (1): 1-12.

Quilis E. (2002). Librería MATLAB de procedimientos de desagregación temporal. INE.

Quilis E.M. (2018). Temporal disaggregation of economic time series: The view from the trenches. Statistica Neerlandica 72 (4): 447-470.

Rao J. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association 91 (434): 499-506.

Rässler S. (2012). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. Vol. 168. Springer Science & Business Media.

Ravikumar P., Liu H., Lafferty J.D. y Wasserman L.A. (2007). SpAM: Sparse Additive Models. En: NIPS. p. 1201-1208.

Reinsch C.H. (1967). Smoothing by spline functions. Numerische mathematik 10 (3): 177-183.

Rey D., Arbués P., López F.A. y Páez A.Using machine learning to identify spatial market segments. A reproducible study of major Spanish markets. Environment and Planning B: Urban Analytics and City Science 0 (0): 23998083231166952. https://doi.org/10.1177/23998083231166952.

Rey-Blanco D., González J. y Sánchez D. (2023a). Relación entre precios de alquiler en portales inmobiliarios y precios de mercado. Evidencias para la Comunidad de Madrid. EURE 0 (0).

Rey-Blanco D., Zofío J.L. y González J. (2023b). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. Expert Systems With Applications 0 (0): 23998083231166952. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121059.

Rico J.R. y Taltavull P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. Expert Systems with Applications 171: 114590.

Robinson P.M. (1988). Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society 931-954.

Rodgers W. (1984). An evaluation of statistical matching. Journal of Business & Economic Statistics 2 (1): 91-102.

Rodriguez López J. (2017). Las viviendas que pudieron hundir la economia española. La caida del mercado de vivienda y sus consecuencias. Cuadernos de Relaciones Laborales 35: 71-99. https://doi.org/10.5209/CRLA.54984.

Rodríguez López J. (2019). El mercado de vivienda mantiene el crecimiento en 2019. Ciudad y Territorio Estudios Territoriales 51 (201): 623-634.

Rodríguez López J. (2022b). Mercado de vivienda y coyuntura económica general. Ciudad y Territorio Estudios Territoriales 54 (214): 1027-1038. https://doi.org/10.37230/CyTET.2022.214.13.

Rodríguez López J. (2022a). El mercado de vivienda resiste las primeras consecuencias de la guerra de Ucrania. Ciudad y Territorio Estudios Territoriales 54 (213): 743-756. https://doi.org/10.37230/CyTET.2022.213.13.

Rodriguez-López J. (2009). Los mercados de vivienda pueden tocar fondo en 2009. Ciudad y Territorio - Estudios Territoriales 41: 365-400.

Rodwin L. (1950). Rent Control and Housing. Social Research 302-319.

Rojo-García J.L. y Sanz-Gómez J.A. (2005). A Bayesian benchmarking method with applications to the Quarterly National Accounts. Luxembourg: Office for Official Publications of the European Communities (ISSN 1725-4825.

Rojo-García J.L. y Sanz-Gómez J.A. (2017). Benchmarking and reconciliation of time series: An applied Bayesian method. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 13 (4): 123.

Román L. álvarez, Muñoz P.A., Barceló C., Brunet J., Azofra L.C., Sáez L.C., Ferreira C., Gálvez J., Gómez M., Rodr’ıguez D.L. y others (2020). El mercado de la vivienda en España entre 2014 y 2019. Documentos ocasionales-Banco de España (13): 1-53.

Romero V., Garmendia M., Ureña Francés J.M. de y others (2014). The Spanish Cadastre: office location, morphologies and dynamics in metropolitan Madrid. Boletıén de la Asociación de Geógrafos Españoles.

Rosen S. (1974). Hedonic prices and implicit markets: product differentiation in

pure competition. Journal of political economy 82 (1): 34-55.

Rosenfeld D. (2022). Using real-time indicators for economic decision-making in government.

Rosenthal L. (1989). Income and price elasticities of demand for owner–occupied housing in the UK: evidence from pooled cross–sectional and time–series data. Applied Economics 21 (6): 761-775.

Roth A.E. (1988). The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press.

Rubin D. (1988). An overview of multiple imputation. En: Proceedings of the survey research methods section of the American statistical association. Citeseer, p. 79-84.

Rubin D. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association 91 (434): 473-489.

Rubin D.B. (1976). Inference and missing data. Biometrika 63 (3): 581-592.

Ruggles N. y Ruggles R. (1974). A strategy for merging and matching microdata sets. En: Annals of Economic and Social Measurement, Volume 3, number 2. NBER, p. 353-371.

Rull J.S. (2018). Consecuencias jurídicas del desistimiento anticipado por parte del arrendatario de un contrato de arrendamiento de inmueble urbano. Revista Crítica de Derecho Inmobiliario 94 (765): 211-235.

Sakia R.M. (1992). The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society: Series D (The Statistician) 41 (2): 169-178.

Samardzhiev K., Fleischmann M., Arribas-Bel D., Calafiore A. y Rowe F. (2022). Functional signatures in Great Britain: A dataset. Data in Brief 43: 108335. https://doi.org/https://doi.org/10.1016/j.dib.2022.108335.

Sánchez-Crespo G. (2002). Introducción a los modelos de superpoblación en las técnicas de muestreo con probabilidades desiguales. Metodología de Encuestas (1): 87-104.

Särndal C.-E. (2007). The calibration approach in survey theory and practice. Survey methodology 33 (2): 99-119.

Särndal C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika 67 (3): 639-650.

Särndal C.-E. y Lundström S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Journal of Official Statistics 24 (2): 167.

Särndal C.-E., Swensson B. y Wretman J. (2003). Model assisted survey sampling. Springer Science & Business Media.

Sax C. y Steiner P. (2013). Temporal Disaggregation of Time Series. The R Journal 5 (2): 80-87. https://doi.org/10.32614/RJ-2013-028.

Sayal H., Aston J.A., Elliott D. y Ombao H. (2017). An introduction to applications of wavelet benchmarking with seasonal adjustment. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180 (3): 863-889.

Scanu M. (2010). Recommendations on statistical matching, Report WP2, ESS-net. Statistical Methodology Project on Integration of Surveys and Administrative Data.

Schafer J. y Graham J. (2002). Missing data: our view of the state of the art. Psychological methods 7 (2): 147.

Schafer J. y Olsen M. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioral research 33 (4): 545-571.

Schapire R.E. (1990). The strength of weak learnability. Machine learning 5 (2): 197-227.

Scher S. y Peßenteiner S. (2021). Temporal disaggregation of spatial rainfall fields with generative adversarial networks. Hydrology and Earth System Sciences 25 (6): 3207-3225.

Schubert E. y Rousseeuw P.J. (2019). Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. En: International conference on similarity search and applications. Springer, p. 171-187.

Schwager S. y Margolin B. (1982). Detection of multivariate normal outliers. The annals of statistics 10 (3): 943-954.

Schwert G.W. (1989). Tests for Unit Roots: A Monte Carlo Investigation. Journal of Business & Economic Statistics 7 (2): 147-159.

Selim H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert systems with Applications 36 (2): 2843-2852.

Seni G. y Elder J. (2010). Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis lectures on data mining and knowledge discovery 2 (1): 1-126.

Shiller R.J. (1991). Arithmetic repeat sales price estimators. Journal of Housing Economics 1 (1): 110-126.

Shiller R.J. (2007b). Low interest rates and high asset prices: An interpretation in terms of changing popular economic models.

Shiller R.J. (2007a). Understanding recent trends in house prices and home ownership. National Bureau of Economic Research.

Shiller R.J. (2008). Derivatives markets for home prices. National Bureau of Economic Research.

Shimizu C., Nishimura K. y Watanabe T. (2016). House prices at different stages of the buying/selling process. Regional Science and Urban Economics 59: 37-53.

Silva J.S. y Cardoso F. (2001). The Chow-Lin method using dynamic models. Economic modelling 18 (2): 269-280.

Similarweb (2022). Clasificación de sitios de internet más visitados de España para diciembre de 2022. https://www.similarweb.com/top-websites/spain.

Simon N., Friedman J., Hastie T. y Tibshirani R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39 (5): 1-13. https://doi.org/10.18637/jss.v039.i05.

Simon W. (2017). Generalized Additive Models: An Introduction with R. 2.ª ed. Chapman; Hall/CRC.

Sirmans S., Macpherson D. y Zietz E. (2005). The composition of hedonic pricing models. Journal of real estate literature 13 (1): 1-44.

Small K.A. y Song S. (1994). Population and employment densities: structure and change. Journal of urban economics 36 (3): 292-313.

Smith L.B. y Tomlinson P. (1981). Rent controls in Ontario: roofs or ceilings? Real Estate Economics 9 (2): 93-114.

Steurer M., Hill R.J. y Pfeifer N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research 38 (2): 99-129.

Stevens J.P. (1984). Outliers and influential data points in regression analysis. Psychological bulletin 95 (2): 334.

Stevenson S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models. Journal of Housing Economics 13 (2): 136-153.

Stewart J.Q. (1947). Empirical mathematical rules concerning the distribution and equilibrium of population. Geographical review 37 (3): 461-485.

Stigler G.J. (1961). Economic problems in measuring changes in productivity. En: Output, input, and productivity measurement. Princeton University Press, p. 47-78.

Stock J.H. (1989). Nonparametric policy analysis. Journal of the American Statistical Association 84 (406): 567-575.

Stone R. (1956). Quantity and Price Indices in National Accounts, esp. Chap.

Strobl C., Boulesteix A.-L., Zeileis A. y Hothorn T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8 (1): 25.

Struijs P. y Daas P. (2014). Quality approaches to big data in official statistics. En: European Conference on Quality in Official Statistics.

Syed I.A. y De Haan J. (2017). Age, time, vintage, and price indexes: measuring the depreciation pattern of houses. Economic Inquiry 55 (1): 580-600.

Tan P.-N., Steinbach M., Karpatne A. y Kumar V. (2018). Introduction to Data Mining (2nd Edition). 2nd ed. Pearson.

Taylor J. y Einbeck J. (2013). Challenging the curse of dimensionality in multivariate local linear regression. Computational Statistics 28: 955-976.

Theil H. (1967). Economics and information theory. North-Holland.

Thériault M., Des Rosiers F. y Joerin F. (2005). Modelling accessibility to urban services using fuzzy logic: A comparative analysis of two methods. Journal of Property Investment & Finance 23 (1): 22-54.

Therneau T., Atkinson B., Ripley B. y Ripley M.B. (2015). Package rpart. Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016).

Thibodeau T.G. (1995). House price indices from the 1984–1992 MSA American Housing Surveys. Journal of Housing Research 439-481.

Tietjen G.L. y Moore R.H. (1972). Some Grubbs-type statistics for the detection of several outliers. Technometrics 14 (3): 583-597.

Tillé Y. y Matei A. (2016). sampling: Survey Sampling. https://CRAN.R-project.org/package=sampling.

Tinsa (2023). Precio de la vivienda en España. https://www.tinsa.es/precio-vivienda/.

Triplett J.E. (1996). The importance of using superlative index numbers. En: CSO Meeting on Chain Indexes for GDP, London, Vol. 26.

Truong Q., Nguyen M., Dang H. y Mei B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science 174: 433-442.

Tucat Pablo L.M. (2021). Poliéticas de oferta para mejorar el acceso a la vivienda de alquiler e España. ESADE. https://www.esade.edu/ecpol/es/publicaciones/politicas-de-oferta-para-mejorar-el-acceso-a-la-vivienda-de-alquiler-en-espan%CC%83a/.

Tukey J.W. (1953). The problem of multiple comparisons. Multiple comparisons.

Turnbull G.K. (1990). The pure theory of household location an axiomatic approach. Journal of Regional Science 30 (4): 549-562.

Uber Inc. (2018). H3: A hexagonal hierarchical geospatial indexing system. https://uber.github.io/h3/#/.

Ulbl M., Verbič M., Lisec A. y Pahor M. (2021). Proposal of real estate mass valuation in Slovenia based on generalised additive modelling approach. Geodetski Vestnik 65 (1).

UNECE (2015). Using administrative and secondary sources for official statistics: A handbook of principles and practices. United Nations Economic Commission for Europe.

Valier A. (2020). Who performs better? AVMs vs hedonic models. Journal of property investment & finance.

Van Buuren S. (2018). Flexible imputation of missing data. CRC press.

Van Der Maaten L., Postma E., Van den Herik J. y others (2009). Dimensionality reduction: a comparative. J Mach Learn Res 10 (66–71).

Vangrevelinghe G. (1966). L'évolution à court terme de la consommation des ménages: connaissance, analyse et prévision. Economie et Statistique 21 (9): 59-102.

Vecchio G. y Martens K. (2021). Accessibility and the Capabilities Approach: a review of the literature and proposal for conceptual advancements. Transport Reviews 41 (6): 833-854.

Verikas A., Lipnickas A. y Malmqvist K. (2002). Selecting neural networks for a committee decision. International Journal of Neural Systems 12 (5): 351-361.

Von Thünen J.H. (1826). El estado aislado.

Waddell P., Berry B.J. y Hoch I. (1993). Residential property values in a multinodal urban area: New evidence on the implicit price of location. The Journal of Real Estate Finance and Economics 7 (2): 117-141.

Wagner C.H. (1982). Simpson's paradox in real life. The American Statistician 36 (1): 46-48.

Wainer H. (1976). Robust statistics: A survey and some prescriptions. Journal of Educational Statistics 1 (4): 285-312.

Wang X., Li K. y Wu J. (2020). House price index based on online listing information: the case of China. Journal of Housing Economics 50: 101715.

Wang X., Wang X. y Wilkes M. (2021). New developments in unsupervised outlier detection. Springer.

Watson G.S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A 359-372.

Wee G.P. van y Vickerman R. (2021). Transport Modes and Accessibility. En: International Encyclopedia of Transportation, Vol. 5. Elsevier.

Wenzlick R. (1952). As I see the fluctuations in the selling prices of single family residences. The Real Estate Analyst 21: 541-548.

Wilkinson R. (1974). The Determinants of Relative House Prices: a case of academic astigmatism? Urban Studies 11 (2): 227-230.

Wingo L. (1961). An economic model of the utilization of urban land for residential purposes. 7 (1): 191-205.

Winter E. (2002). The shapley value. Handbook of game theory with economic applications 3: 2025-2054.

Witte A.D., Sumka H.J. y Erekson H. (1979). An estimate of a structural hedonic price model of the housing market: an application of Rosen's theory of implicit markets. Econometrica: Journal of the Econometric Society 1151-1173.

Wong D.W. (2004). The modifiable areal unit problem (MAUP). WorldMinds:

geographical perspectives on 100 problems: commemorating the 100th anniversary of the association of American geographers 1904–2004 571-575.

Wood S.N., Goude Y. y Shaw S. (2015). Generalized additive models for large data sets. Journal of the Royal Statistical Society: Series C (Applied Statistics) 64 (1): 139-155.

Wood S.N., Li Z., Shaddick G. y Augustin N.H. (2017). Generalized additive models for gigadata: modeling the UK black smoke network daily data. Journal of the American Statistical Association 112 (519): 1199-1210.

Worzala E., Lenk M. y Silva A. (1995). An exploration of neural networks and its application to real estate valuation. Journal of Real Estate Research 10 (2): 185-201.

Wright M.N. y Ziegler A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Wu C. y Sitter R. (2001). A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association 96 (453): 185-193.

Wu Y., Wei Y.D. y Li H. (2020). Analyzing spatial heterogeneity of housing prices using large datasets. Applied Spatial Analysis and Policy 13 (1): 223-256.

Wyngarden H. (1927). An Index of Local Real Estate Prices. University of Michigan, School of business administration.

Xiao Y., Chen X., Li Q., Yu X., Chen J. y Guo J. (2017). Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data. ISPRS International Journal of Geo-Information 6 (11): 358.

Yang Y. (2007). Consistency of cross validation for comparing regression procedures. The Annals of Statistics 35 (6): 2450-2473.

Yatchew A. (1997). An elementary estimator of the partial linear model. Economics letters 57 (2): 135-143.

Zaier L.H. y Abed M. (2014). Temporal Disaggregation of Economic Time Series using Artificial Neural Networks. Communications in Statistics-Theory and Methods 43 (8): 1824-1833.

Zani S. (1970). Sui criteri di calcolo dei valori trimestrali di tendenza degli aggregati di contabilità nazionale. Studi e Richerche, Facoltà de Economia e Commercio, Università degli Studi di Parma 7: 285-349.

Zhang G. y Lu Y. (2012). Bias-corrected random forests in regression. Journal of Applied Statistics 39: 151-160. https://doi.org/10.1080/02664763.2011.578621.

Zhang L.C. y Nguyen N. (2020). An appraisal of common reweighting methods for nonresponse in household surveys based on Norwegian Labour Force Survey and Statistics on Income and Living Conditions Survey. Journal of Official

Statistics 36 (1): 151-172.

Zhou Z.-H. (2021). Ensemble learning. En: Machine learning. Springer, p. 181-210.

Zhu L. y Zhang H. (2021). Analysis of the diffusion effect of urban housing prices in China based on the spatial-temporal model. Cities 109: 103015.

Zhu Y. y Bradic J. (2017). Breaking the curse of dimensionality in regression. arXiv preprint arXiv:1708.00430.

Zimmerman D.W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. The Journal of Experimental Education 64 (1): 71-78.

Zimmerman D.W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. The Journal of experimental education 67 (1): 55-68.

Zulkifley N.H., Rahman S.A., Ubaidullah N.H. y Ibrahim I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education & Computer Science 12 (6).

Zyga J. (2019). Data selection as the basis for better value modelling. Real Estate Management and Valuation 27 (1): 25-34.