

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.

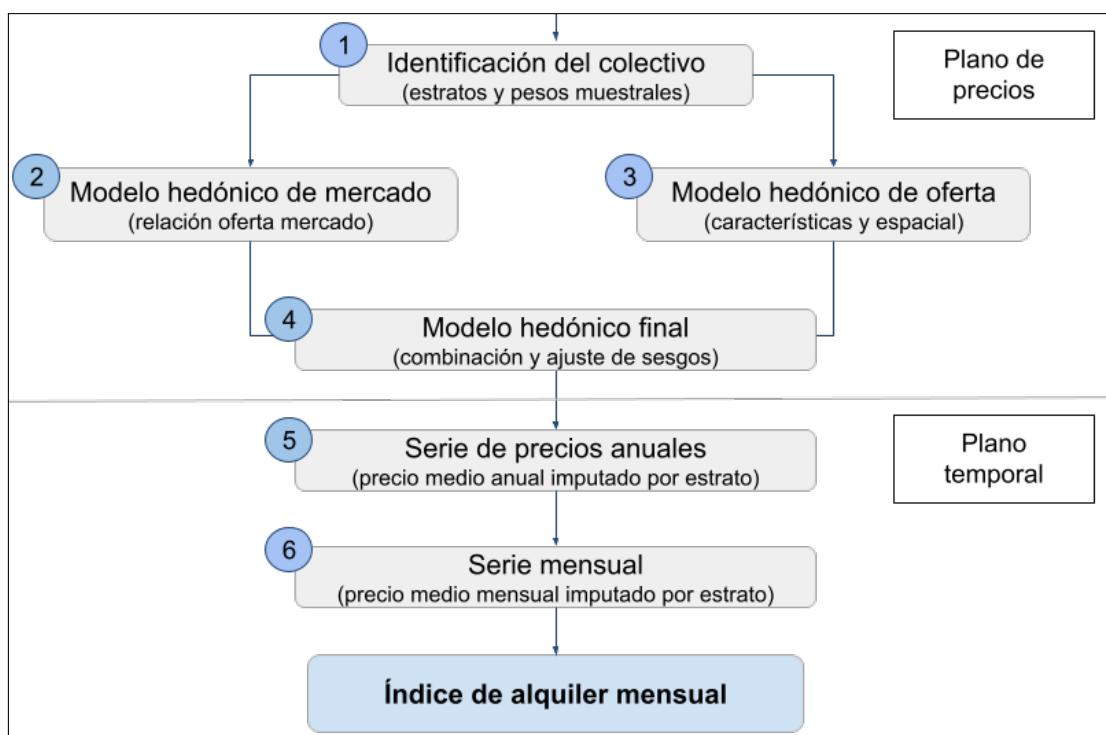
Sin embargo, en 2020, el MITMA comenzó a publicar el índice de precios del alquiler, que ofrece estadísticas estatales de precios por metro cuadrado procedentes de registros tributarios. Adicionalmente, existen algunos registros regionales y locales, como el caso de INCASOL¹ en Cataluña, que no son abiertos 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

¹INCASOL, 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, ponderación y calibración de muestra	Construcción del colectivo y cálculo de los elevadores muestrales sobre la población de oferta	Precios
	Creación de modelos hedónicos de precios de alquiler	Precios
Modelos de valoración por precios hedónicos	Creación de modelos hedónicos de precios de oferta	Precios
	Creación de modelo de enlace entre 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.

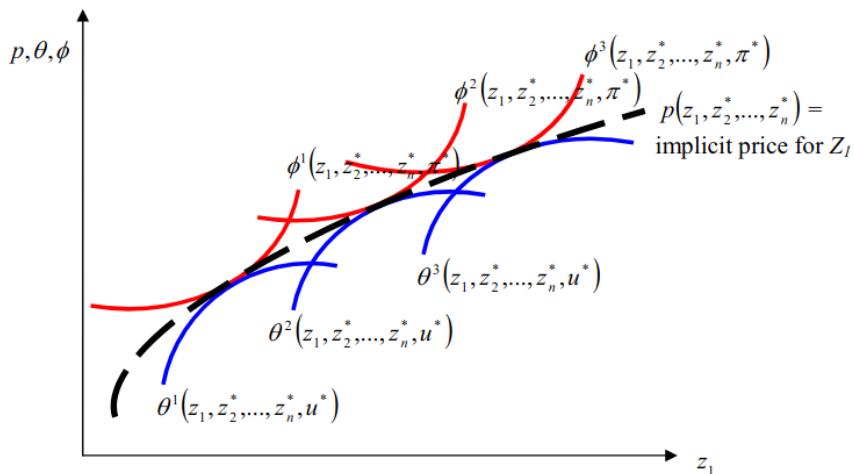
La función de puja se puede escribir como:

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

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:

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

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

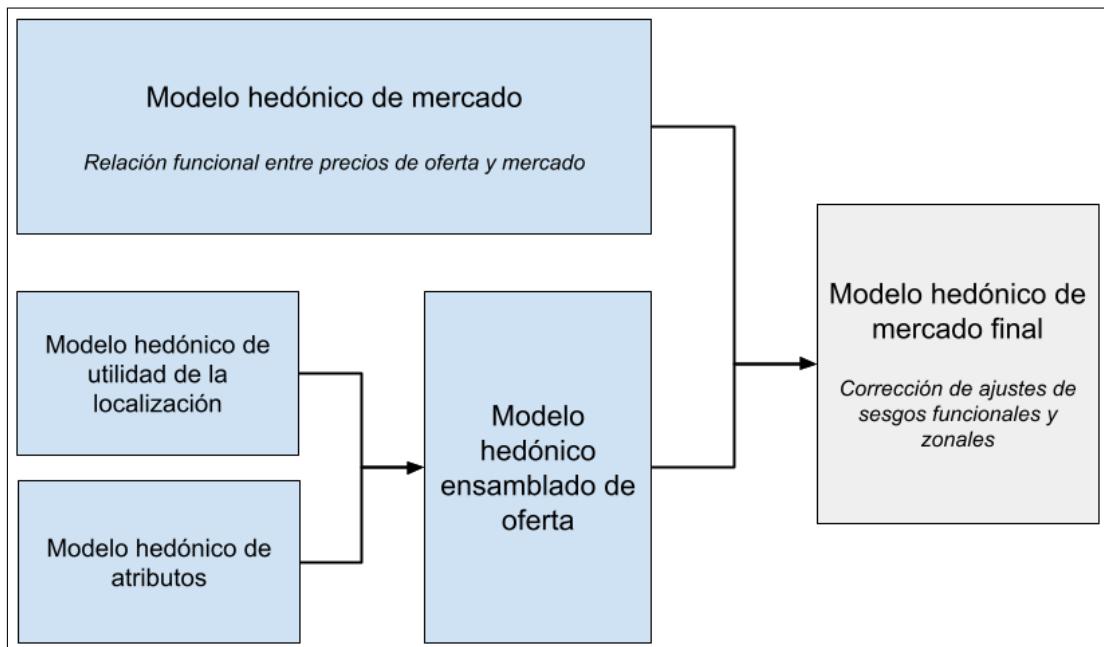
Fuente: Des Rosiers y Thériault (2006).

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 estadístico². 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

²En 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], semilogarítmica [2.4] y logarítmica [2.5].

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

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).

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

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

possible aplicar logaritmos a estas variables³ y, por tanto, dichas variables se modelan en forma semilogarítmica (Bover y Velilla, 2001).

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

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:

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

donde D_{nk}^t son variables ficticias dicótomicas que representan el tiempo⁴; 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

³Cuando la variable toma el valor cero.

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

⁵Estas 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 .

⁶Por 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 (WLS⁷). 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

⁷Del 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 splines⁸ (Reinsch, 1967); el método basado en los vecinos más cercanos (kNN⁹) (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”. Esta actividad está regulada en España por la normativa ECO/805/2003¹⁰.

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.

⁸Una 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”.

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

¹¹También 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, estadístico¹² o aprendizaje de máquinas¹³, 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 algoritmos¹⁴, 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 es sensiblemente menor que los que no lo

¹²Dependiendo del tipo de técnica aplicada, en particular aquellas en las que existe base estadística en el proceso, algunos autores lo califican como aprendizaje estadístico.

¹³También referido en su original en inglés “*machine learning*”.

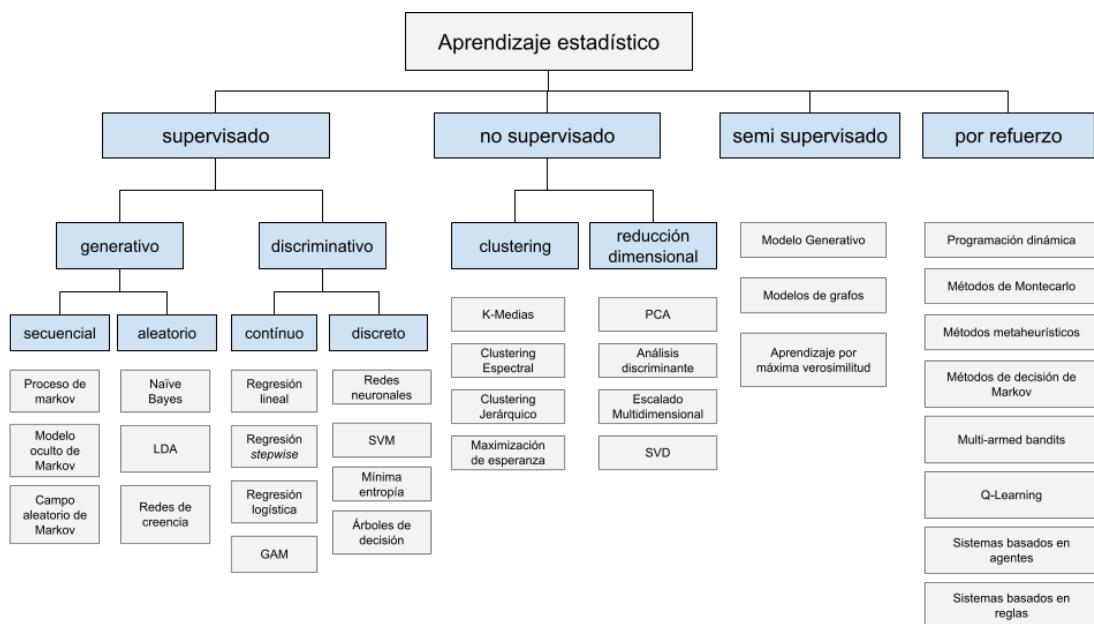
¹⁴La complejidad computacional se puede definir como el nivel de exigencia de recursos: proceso, memoria o tiempo, necesarios para resolver un problema con un programa de ordenador.

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 ANN¹⁵) fue propuesto por Pitts y McCulloch (1943) en el artículo titulado “*Un cálculo lógico de*

¹⁵Red 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 servicios 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 inefficientes 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 vecinas¹⁶. 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 extensible a otras áreas con patrones urbanos y de mercado diferentes. Esta cuestión puede ser especialmente problemática

¹⁶Un 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 de 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 (<i>spatial lag</i>) 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 <i>et al.</i> , 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 <i>et al.</i> , 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 <i>et al.</i> , 2003).
	Partición espacial	Dividen el área de estudio en subregiones homogéneas y estiman modelos de regresión hedónica separados para cada subregión (Kim <i>et al.</i> , 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

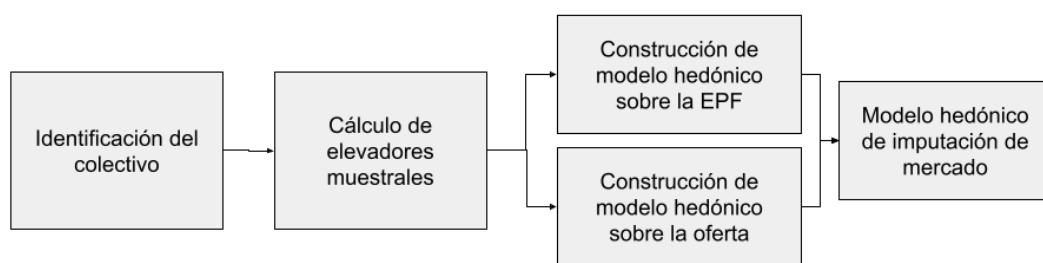
Tabla 2.2. Taxonomía de métodos hedónicos geográficos

2.2.2 Modelo hedónico de mercado

El primer paso en la construcción del modelo de mercado es la correcta caracterización del colectivo de mercado¹⁷, 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 el 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

¹⁷Todas las viviendas que se encuentran en régimen de alquiler.

¹⁸La 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).

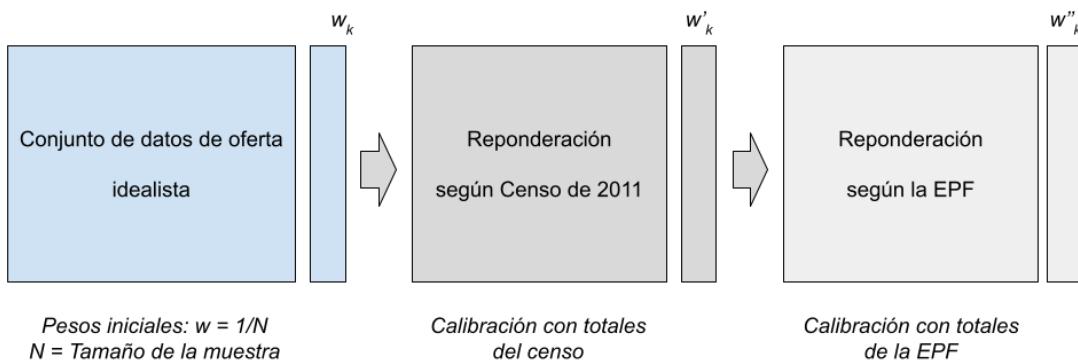
- Cada celda de la rejilla debe disponer de información suficiente, que en nuestro caso se ha establecido en un mínimo de 30 anuncios por celda.

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 vivienda²⁰, *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 autónoma²¹, *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:

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

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.

¹⁹El *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.

²⁰Para el modelo base de oferta se utilizaría el término \hat{P}_o .

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

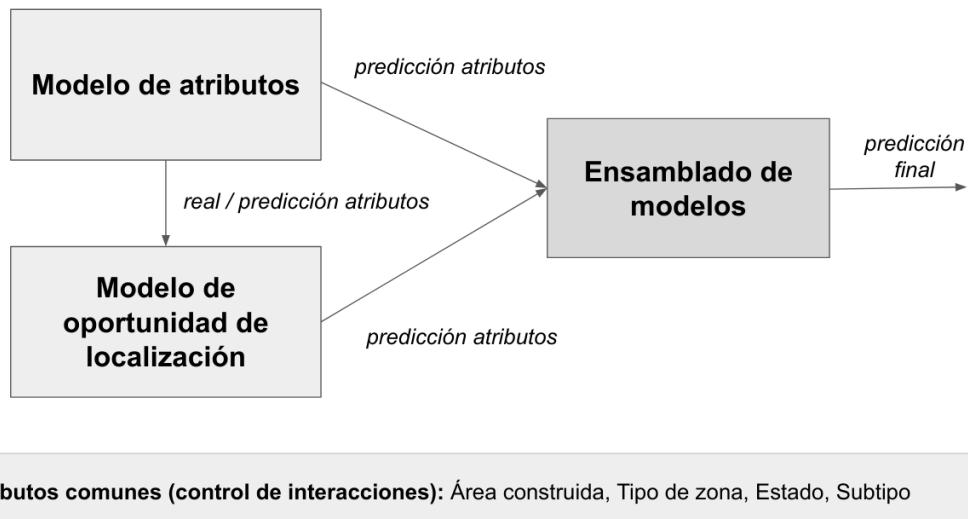
2.2.3 Modelo hedónico de oferta

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.

²²Random 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

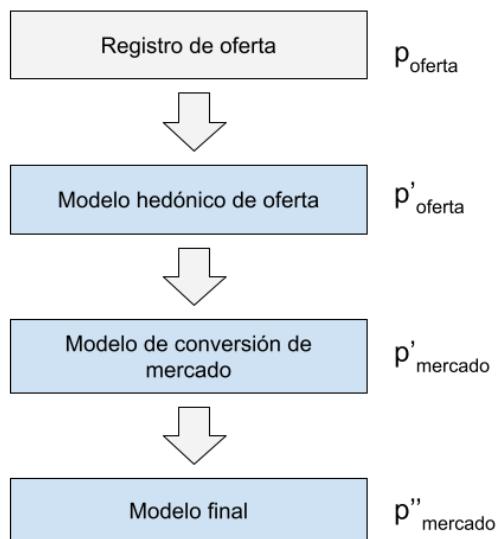
El modelo de localización anterior no utiliza la ubicación geográfica del inmueble, sino que utiliza la utilidad asociada a dicha ubicación. El motivo procede de la justificación del propio modelado hedónico, dado que el precio se explica, parcialmente, por la prima que están dispuestos a pagar los usuarios por la utilidad situacional. Esta utilidad se puede expresar, numéricamente, en términos de accesibilidad²³ 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).

²³Véase 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 erráticos²⁴.

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).

²⁴Principalmente 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 construcciones 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 transacciones²⁵, 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.

²⁵La 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:

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

- 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 períodos comparados. En este caso las cantidades de cada estrato i se fijan de antemano y no proceden de ningún período (q_i). Este tipo de índice fue propuesto por primera vez por Lowe (1824), y se calcula según la siguiente expresión:

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

- 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:

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

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

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

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 ($q_{t0} \dots q_{t_n}$), mientras que el índice de Paasche mide el valor de la cesta ($q_{t0} \dots q_{nt}$).

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

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

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).

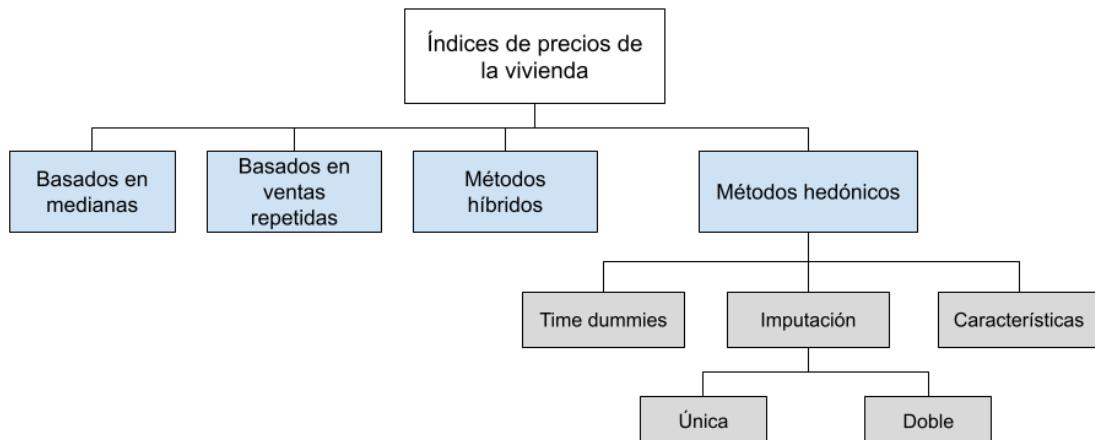
Los métodos mencionados anteriormente se basan en ratios y productos entre precios y cantidades, muchos de ellos están recogidos en el trabajo de Fisher (1922a). No obstante, existen alternativas como los índices basados en diferencias (Diewert, 2005). Esta línea iniciada por Bennet (1920) y Montgomery (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 medianas, 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:

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

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

donde $\text{med}(P_t^e)$ es la mediana del precio de la vivienda para un estrato de la muestra e , en el momento del tiempo t . Mientras que $\text{med}(P_0^e)$ 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:

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

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:

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

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:

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

Finalmente, el índice se construye como:

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

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

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

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:

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

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:

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

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.

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

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 $\gamma_n, n = 1, \dots, T$, 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 inmobiliarias (MLS²⁷) 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

²⁷MLS, acrónimo de “Multiple Listing Services”, se refiere a las asociaciones entre agencias inmobiliarias estadounidenses.

comportamiento del índice Case-Shiller²⁸, 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 procede 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 oficialmente²⁹). 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 Trends³⁰ 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.

²⁸El 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.

²⁹Datos estadísticos basados en registros oficiales (OMI) del Ministerio de Hacienda Italiano.

³⁰Google 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.

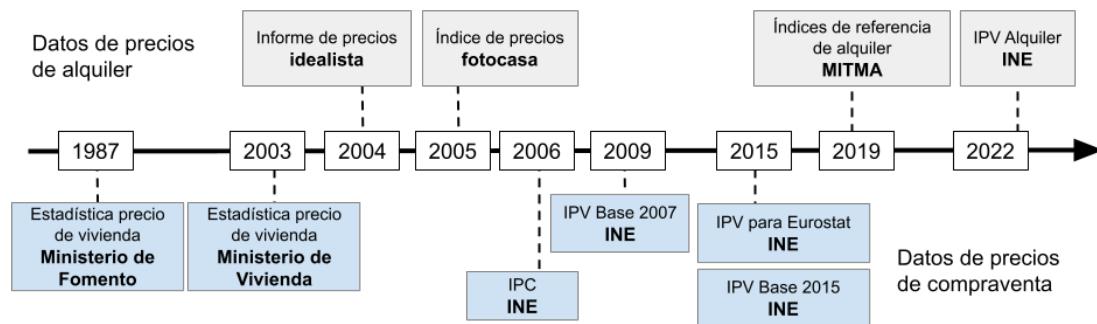
Posteriormente, en 2015, el INE actualizó la metodología para integrarse en el índice de precios europeo armonizado de la vivienda de Eurostat, momento en el que actualizó la metodología anterior del IPV. En 2019, el Ministerio de Transportes, Movilidad y Agenda Urbana publicó la primera base de datos del alquiler desglosada de ámbito nacional³¹. 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.

³¹Excluyendo 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 / y 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 IRPF³².

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

³²Impuesto 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 libre³³ 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 HPI³⁴, 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 ...”

³³Viviendas residenciales que no son ni viviendas sociales o de protección oficial (VPO).

³⁴Armonized 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 IPC³⁵, 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 provincia³⁶. La última modificación de la metodología corresponde a 2017³⁷ 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 encuestados.

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 Similarweb³⁸, 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 venta³⁹, mientras que Idealista contabilizó 98.987 viviendas en alquiler y 684.073 en venta⁴⁰.

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

³⁵Índice de Precios al consumo

³⁶El dato del IPC armonizado se extrae filtrando el subgrupo ECOICOP del alquiler de la vivienda (041).

³⁷Aunque 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).

³⁸Similarweb (2022) es una compañía que ofrece servicios de análisis web, como medidas de volumen de tráfico y de usuarios a sitios de internet.

³⁹Datos de anuncios publicados en Idealista (Idealista, 2021) el 12 de septiembre de 2021.

⁴⁰Datos 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 OCDE⁴¹ (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

⁴¹Organización para la Cooperación y el Desarrollo Económicos.

⁴²La 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. Esta 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 Denton⁴³.

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:

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

donde la verosimilitud se calcula es el producto de probabilidades de un conjunto de parámetros de calidad⁴⁴ c para la serie desagregada H con el método m . Con el objeto de facilitar el cálculo, se asume independencia de los sucesos.

⁴³El método Denton se describe con detalle en el Anexo 7a del capítulo 7.

⁴⁴Son 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-Trager⁴⁵.

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:

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

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.

⁴⁵Estos 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 Familiares⁴⁶ 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.

⁴⁶Más 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 dígitos⁴⁷), 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 afijación⁴⁸, 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 excede 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/m ²	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 la distribución, eliminando todos aquellos valores más allá de 1.5 veces el rango intercuartílico.

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 Viviendas⁴⁹ 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

Capítulo 2. Metodología y fuentes de información

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 Madrid⁵⁰, 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

⁵⁰Accesible a través de su web en <http://www.idealista.com>

Capítulo 2. Metodología y fuentes de información

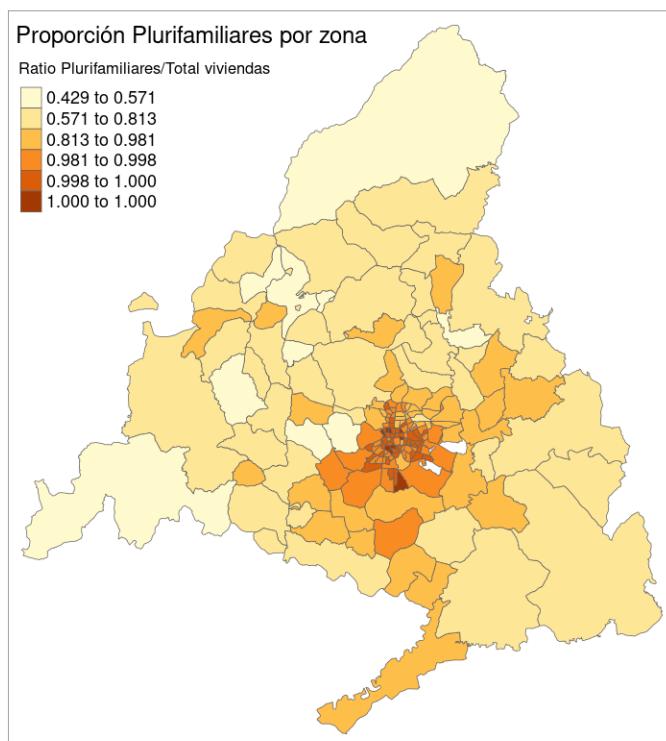
MAXBUILDINGFLOOR	Número de pisos en edificio	Edificio
BATHNUMBER	Número de baños	Estructura
BEDROOMNUMBER	Número de dormitorios	Estructura
CONSTRUCTEDAREA	Superficie total en metros cuadrados	Estructura
ENERGYCERTIFICATIONID	Código de certificado energético	Estructura
FLATLOCATION	Indica si el piso es 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
HASKINGSPACE	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 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 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

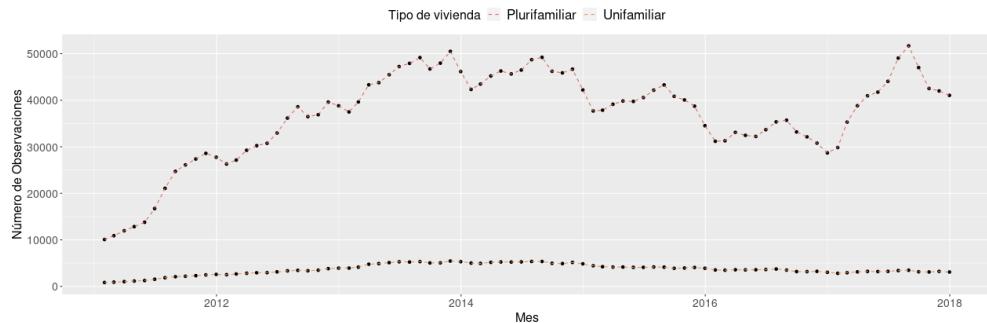


Fuente: elaboración propia.

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 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

⁵¹Organización 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 GeoFabrik⁵², 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 transporte⁵³, 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



El dato utilizado corresponde a una extracción de la Sede Electrónica del Catastro⁵⁴ 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 principal⁵⁵ 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 centroide⁵⁶ 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 destino⁵⁷.

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>

⁵⁵Se corresponde a los metros construidos para uso destino mayoritario de la finca catastral.

⁵⁶Centro geométrico de un polígono.

⁵⁷El 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
Construcción	3.050.967	4.270.251	7.321.218
Construcción 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, donde 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 usuarios⁵⁸. 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 “[...]

⁵⁸Por 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 Hawkins (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 I⁶⁰ y tipo II⁶¹; 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

⁵⁹La 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.

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

⁶¹El 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).

Cousineau (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 en 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).

⁶²Esta 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

⁶³En 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 límites 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 intercuartílico. Para el caso de la barrera cercana los límites se calcularían con la siguiente expresión analítica:

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

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 más 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 períodos. 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

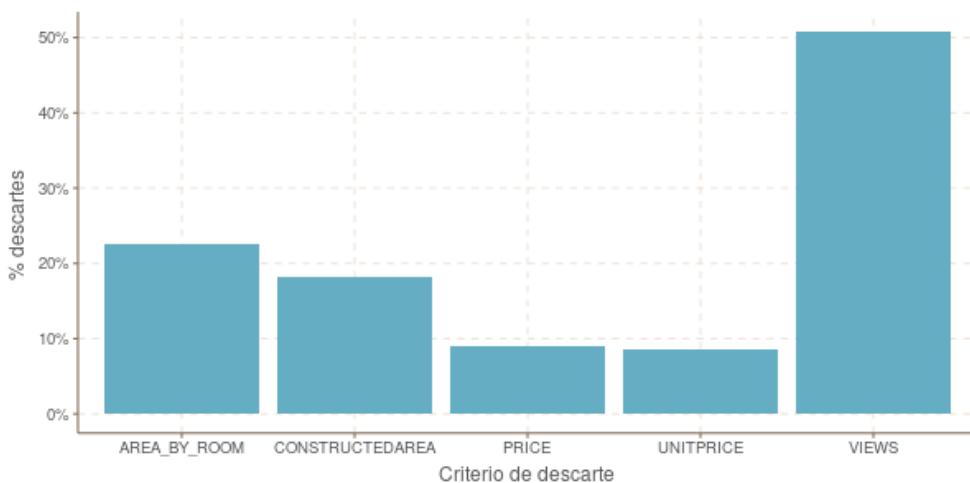
⁶⁴Es importante destacar que la extracción se realiza sobre la base de Idealista y no es una captura

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 española⁶⁷ 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

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

⁶⁶Un 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.

⁶⁷Orden 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 mínimos 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).

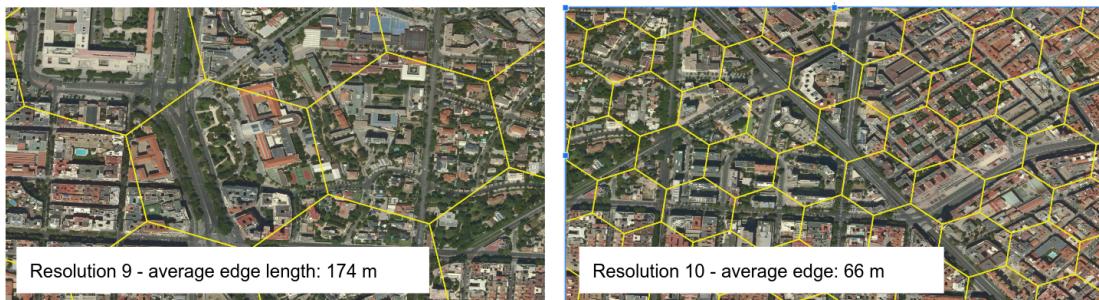
Además de los métodos paramétricos, existen métodos paramétricos que no asumen una distribución en la presencia de valores ausentes, como los métodos apoyados en los K-Vecinos más próximos (Fix y Hodges, 1989), o los basados en imputación en modelos de aprendizaje automático. Ambas familias permiten 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



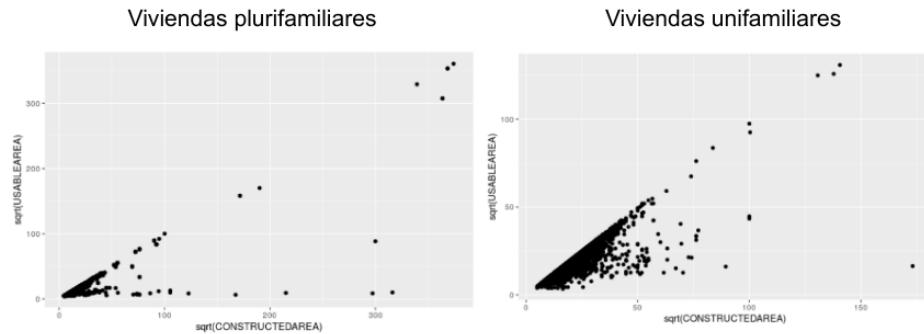
Fuente: elaboración propia.

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 CART⁶⁸ (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.

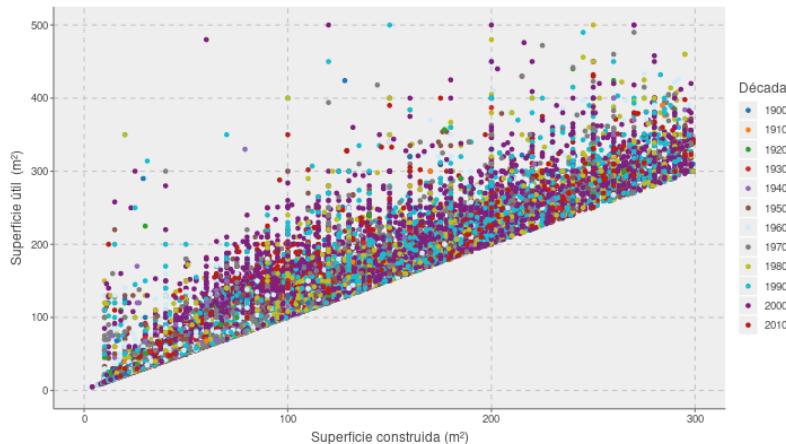
Figura 2.18. Relación entre área útil y área construida



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:

⁶⁸En 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 commensurabilidad 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.

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

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.

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

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.