Capítulo 3

Modelo de mercado

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

— George E.P. Box

3.1 Introducción

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

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

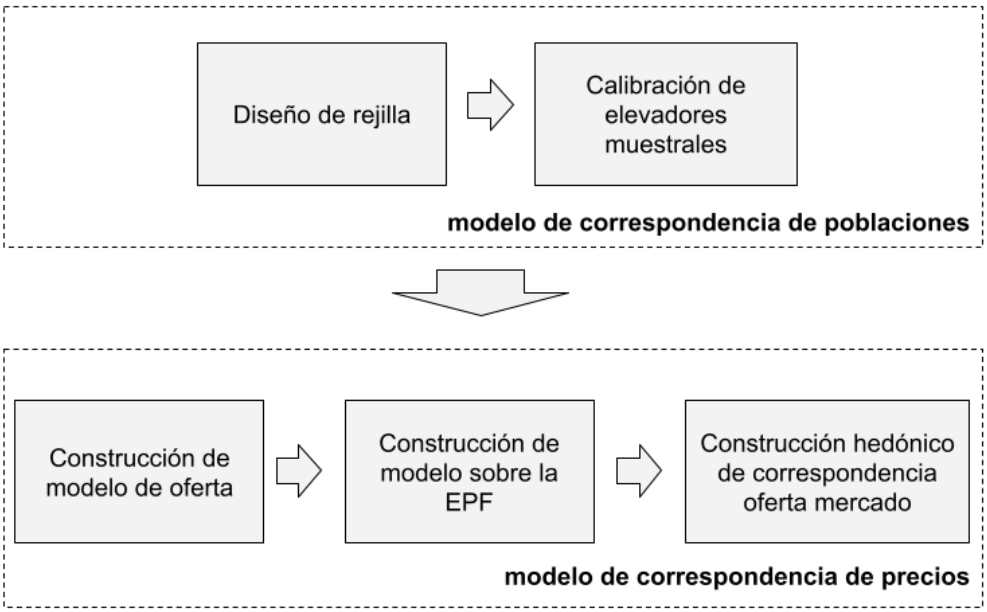
Por otra parte, es habitual que los registros de portales inmobiliarios cuenten con duplicidades o fenómenos de sobre e infrarrepresentación para ciertos segmentos del mercado (Loberto *et al.*, 2018; Pangallo y Loberto, 2018). Algunos de ellos, motivados por la diversidad de patrones de comportamiento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

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

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

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

Figura 3.1. Flujo de trabajo del modelo de mercado



Fuente: elaboración propia.

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

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

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

3.2 Modelo de correspondencia de poblaciones

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

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

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

3.2.1 Métodos de correspondencia estadística

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

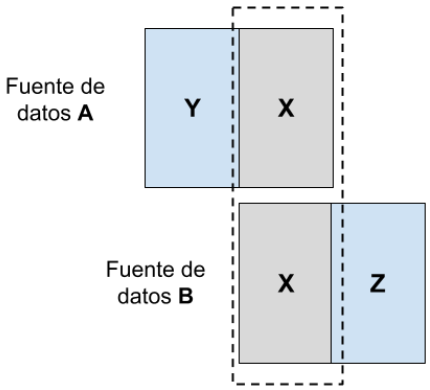
[7] En este caso se refiere a ellas como variables sombra (*shadow*) en lugar de "proxy".

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

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

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

Figura 3.2. Independencia condicional



Fuente: elaboración propia.

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

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

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

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

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

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

3.2.1.1 Métodos y medidas de distancia

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

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

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

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

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

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

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

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

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

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

3.2.2 Diseño de la muestra

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

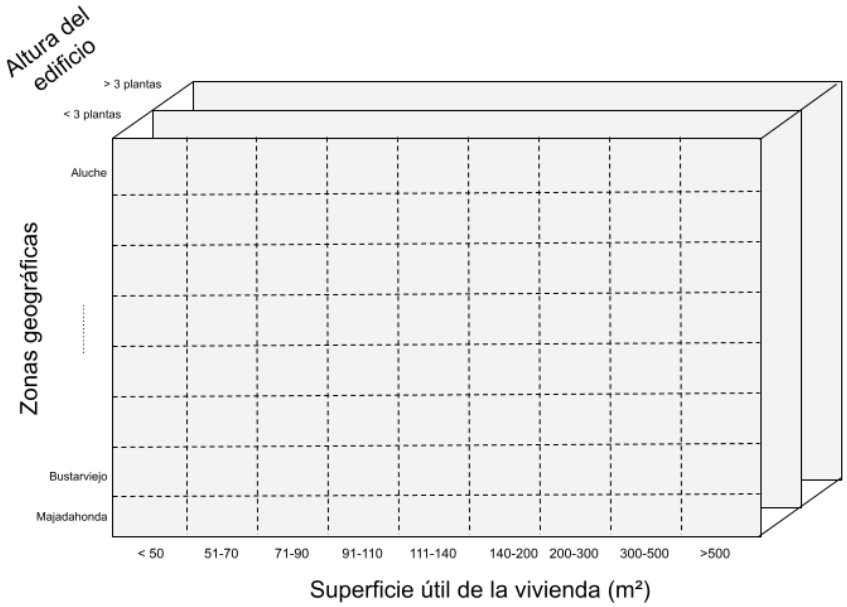
Para lograr una explotación detallada del índice de precios, se realizará una estratificación basada en el censo de viviendas que se aplicará sobre la base de datos de oferta. Como no se dispone del dato de censo a partir del segundo año y siguientes, se hará un ajuste de los pesos poblacionales del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

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

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

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

Figura 3.3. Ejemplo de rejilla para una población de viviendas



Fuente: elaboración propia.

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

3.2.2.1 Diseño de rejillas los procesos de calibración

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

[11] Véase apartado 2.4.6.2

los dominios de interés del análisis.

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

Tabla 3.1. Resumen de variables de cada rejilla de trabajo

Variable	**Descripción**	**Niveles distintos**	**Censo**	**EPF**
ANOCONSC	Año construcción	4	X	X
ASCENSOR	Ascensor	2	X	
GARAJE	Garaje	2	X	
LOCATION	Zona	178	X	
NHAB	Número Habitaciones	4	X	X
PLANTAS	Plantas	5	X	X
SUT	Superficie Útil	4	X	X
CAPROV	Código de provincia	1		X
DENSI	Densidad de población	3		X
FACTORGASTOT6	Factor de gasto	3		X
INTERINPSP	Ingresos netos	3		X
TAMAU	Población del municipio	5		X
TIPOCASA	Tipo de vivienda	3		X
TIPOEDIF	Tipo de edificio	4		X
ZONARES	Tipo de zona	3		X

Fuente: elaboración propia

Las variables utilizadas para la rejilla del Censo usan la misma nomenclatura que las variables de la fuente original (para más detalle véase epígrafe 2.4.2), y son las siguientes:

- *CAPROV*: Es capital de provincia o no.
- *LOCATION*: Código de Barrio en el caso de Madrid y Código de municipio (o grupos de municipios) para el resto.
- *SUT*: Superficie útil (4 niveles).
- *NHAB*: Número de habitaciones (4 niveles).

- *PLANTAS*: Número de plantas del edificio (5 niveles).
- *ASCENSOR*: Tiene ascensor (sí y no).
- *GARAJE*: Tiene garaje (sí y no).
- *ANOCONSC*: Año de construcción. Se usan 4 niveles, anteriores y hasta 1970, de 1971 a 1980, de 1981 a 1990, de 1991 a 2000 y de 2001 en adelante.

Para la calibración según la EPF se utiliza un número mayor de variables, que se enumeran a continuación[12]:

- *CAPROV*: Código de provincia (en este caso esta variable no es relevante porque solo se trabaja en el ámbito de la provincia de Madrid).
- *SUT*: Superficie útil (4 niveles), es importante tener en cuenta que este campo en la EFP viene limitado al intervalo entre 45 y 300 m^2.
- *NHAB*: Número de habitaciones (4 niveles).
- *ANOCONSC*: Año de construcción (4 niveles).
- *TIPOEDIF*: Tipo de edificio (4 niveles).
- *TIPOCASA*: Tipo de vivienda (3 niveles).
- *ZONARES*: Tipo de zona residencial (3 niveles). Se simplifica la clasificación original de la EPF en tres grados, zona de renta baja, media y alta.
- *INTERINPSP*: Intervalo de ingresos mensuales netos totales de cada miembro del hogar (ver definición en la descripción de la fuente de datos EPF). En este caso lo simplificamos a 3 niveles, renta baja (2 primeros niveles en la EPF), media (segundo y tercer nivel) o alta (tres niveles más altos de la variable en la EPF).
- *FACTORGASTOT6*: Gastos familiares, utilizando la codificación de la variable *factorGASTOT6* de la EPF. En este caso, se recodifican los valores originales de la EPF para tener 3 niveles: el primero, gasto bajo-medio que recoge los 4 primeros niveles iniciales; el segundo, gasto medio alto del quinto original; y el tercero, el gasto alto del sexto original.
- *DENSI*: Densidad de población del municipio (3 niveles).
- *TAMAMU*: Tamaño del municipio en población (5 niveles).

Dado que los datos de ingresos por miembro del hogar *INTERINPSP* y el factor de gasto *factorGASTOT6* son de utilidad para el enlace pero no se encuentran en el conjunto de oferta, se añaden en este último con un proceso de imputación múltiple no paramétrico[13] basado en modelos.

El modelo de imputación del factor de gasto se estima según las características de la vivienda y de la zona, a partir de los microdatos de la EPF para toda España. La forma funcional del modelo se define según la siguiente expresión analítica:

[12] El detalle de cada una de las variables de la EPF se describe en el epígrafe 2.4.1

[13] Se utiliza un modelo del tipo *Random Forests* mediante la librería *ranger* de R (Wright y Ziegler, 2015), para más información véase el Anexo 3b.

$$factorGASTOT6 \leftarrow TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + SUPERF + ANNOCON + DENSI + INTERINPSP + NHABIT + CCAA + CAPROV \quad [3.2]$$

donde $CCAA$ representa el código de comunidad autónoma. La configuración de este modelo utiliza 1.000 árboles, con un parámetro $mtry$[14] es igual a 10, y utiliza como pesos el factor de elevación de cada registro del fichero de la EPF.

Para el caso del modelo de ingresos, se utiliza el fichero de la renta media *per cápita* por sección censal de la Comunidad de Madrid. Al disponerse solamente de datos desde el 2016 en adelante, para los años 2015 y anteriores se asignan los ingresos del 2016.

3.2.2.2 Tratamiento de zonas geográficas

La segmentación zonal de la rejilla utilizada en la calibración censal trabaja con dos tipos de áreas[15]: la ciudad de Madrid y el resto de la provincia. En la ciudad, se toma como tamaño de área de trabajo el barrio, puesto que la sección censal no ofrece soporte suficiente de datos en oferta. Para el resto de la provincia se utiliza el municipio. En los casos en los que las celdas no cuentan con observaciones suficientes[16] se agrupan en zonas de orden superior según criterios de similitud inmobiliaria . El número total de zonas es 178, sumados barrios y áreas municipales .

Para la capital, los 128 barrios de la ciudad se agrupan finalmente en 112 zonas de trabajo, porque, como se ha indicado anteriormente, algunos no contaban con suficientes registros. Cada agrupación de dos barrios requiere cumplir el criterio de ser adyacentes geográficamente, y similares en características inmobiliarias y demográficas. Las zonas agregadas se denominan con el literal compuesto de los nombres de los barrios originales (ejemplo “El Pardo - Mirasierra”). Las zonas que se han unificado son:

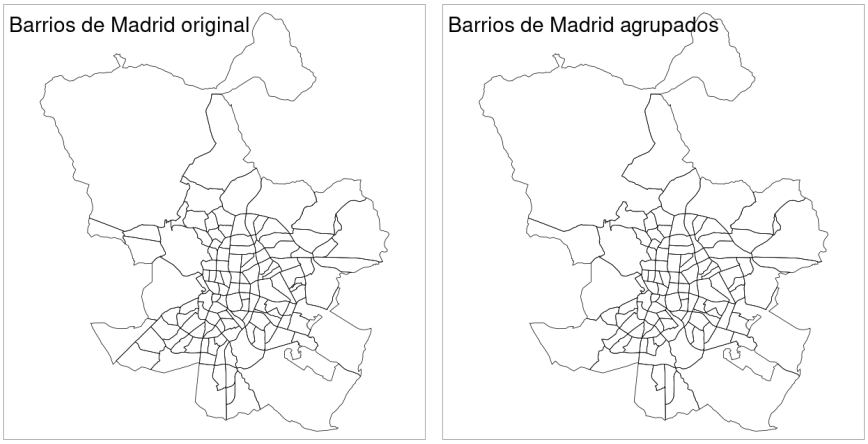
- Códigos 24 y 27: Legazpi - Atocha.
- Códigos 81 y 87: El Pardo - Mirasierra.
- Códigos 82 y 83: Fuentelarreina - Peña Grande.
- Códigos 104, 105, 106 y 107: Aluche - Campamento - Cuatro Vientos - Las Águilas.

[14] Indica el número de variables sobre las que se hacen los cortes del árbol de decisión.
[15] La selección de estos dos niveles está condicionada a las limitaciones de definición de detalle geográfico del Censo de Viviendas.
[16] Se ha tomado como requisito que cada celda geográfica debe contener al menos 30 registros por cada periodo anual.

- Códigos 121, 122 y 123: Orcasitas - Orcasur - San Fermín.
- Códigos 141 y 142: Pavones - Horcajo.
- Códigos 157 y 158: Colina - Atalaya.
- Códigos 161 y 162: Palomas - Piovera.
- Códigos 172 y 174: San Cristóbal - Los Rosales.
- Códigos 202 y 203: Hellín - Amposta.
- Códigos 211 y 212: Alameda de Osuna - Aeropuerto.

En la Figura 3.4, se muestran la división original y la final agrupada, se aprecia que no hay alteraciones sustanciales en la distribución geográfica.

Figura 3.4. Barrios de Madrid originales y agregados



Fuente: elaboración propia.

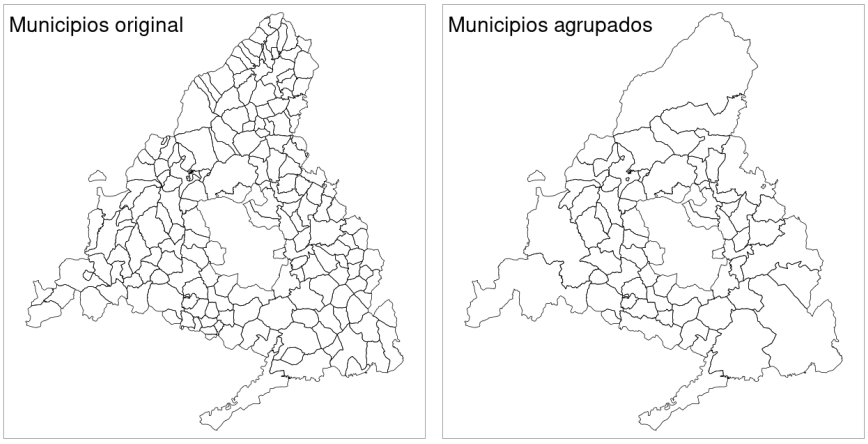
En el resto de la Comunidad de Madrid es más frecuente encontrar zonas de poco soporte, las cuales se han agrupado en municipios con un criterio parecido de similitud. Aunque en este caso, no se exige que el barrio sea adyacente sino que pertenezca a la misma área geográfica dentro de la Comunidad de Madrid (por ejemplo corredor del Henares, o municipios de la sierra de Guadarrama). Se crean las 22 nuevas zonas agrupadas:

- Área de Área de Manzanares el Real.
- Área de Área de Fuente el Saz de Jarama.
- Área de Área de la Sierra norte.
- Área de Daganzo de Arriba - Ajalvir.
- Área de la Cabrera - Torrelaguna.
- Área de Pedrezuela.
- Área de Guadalix de la Sierra.
- Guadarrama - Alpedrete.
- Área de Navacerrada.
- Área de Collado Mediano.

- Área de Mataelpino/Cerceda - Moralzarzal.
- Villanueva del Pardillo - Villanueva de la Cañada.
- Área de Brunete - Quijorna.
- Área de San Martín de Valdeiglesias - Cadalso de los Vidrios - Villa del Prado-Navas del Rey.
- Área de Colmenar del Arroyo.
- Área de Humanes.
- Griñón - Cubas de la Sagra - Casarrubuelos - Batres - Serranillos del Valle - Torrejón de Velasco.
- Área de Perales de Tajuña - Nuevo Baztán - Villarejo.
- Área de Colmenar de Oreja - Chinchón.
- Área de Villalbilla - Loeches.
- Área Meco.
- Área de Mejorada.

En este caso, como expresa la Figura 3.5, es necesario agrupar gran parte de los municipios rurales con menor población. Se parte de 178 municipios, que se consolidan en 66 zonas.

Figura 3.5. Municipios de la Comunidad de Madrid originales y agregados



Fuente: elaboración propia.

3.2.3 Reponderación de los elevadores muestrales

El proceso de correspondencia estadística que relaciona las poblaciones de oferta y la de mercado se realiza mediante un proceso de cálculo de elevadores muestrales de la oferta[17], por medio de un proceso de calibración. Dicho proceso de transformación de los pesos muestrales se denominará reponderación.

La reponderación es una técnica estadística usada comúnmente para compensar los errores de falta de respuesta y cobertura de una muestra. Como caso particular de ella, la calibración es un método que permite estimar los pesos adecuados para que la muestra reproduzca los totales de la población de estudio. Se utiliza habitualmente en muestreo de encuestas (Lohr, 2019), y mejora la calidad de la información mediante el uso de datos auxiliares, corrige de sesgos, y asegura el cumplimiento de suficiencia estadística de Fisher (Kullback, 2012) en la muestra de trabajo.

La falta de respuesta es uno de los principales problemas de cualquier proceso de muestreo, y es común encontrarla cuando se trabajan con datos de portales inmobiliarios (Bricongne *et al.*, 2023; Loberto *et al.*, 2018; Pangallo y Loberto, 2018). Lohr (2019) define el fenómeno de falta de respuesta unitaria como la ausencia de un registro completo, en nuestro caso, por ejemplo, la inexistencia de alquiler social en la muestra o la ausencia de registros de un tipo de vivienda en un estrato. Este fenómeno no debe confundirse con la presencia de valores no observados en los conjuntos de datos (como la ausencia del alguna variable en algunas de las observaciones). Los efectos de ignorar la falta de respuesta puede implicar serios problemas en los resultados del estudio (Fuller, 2011; Lohr, 2019).

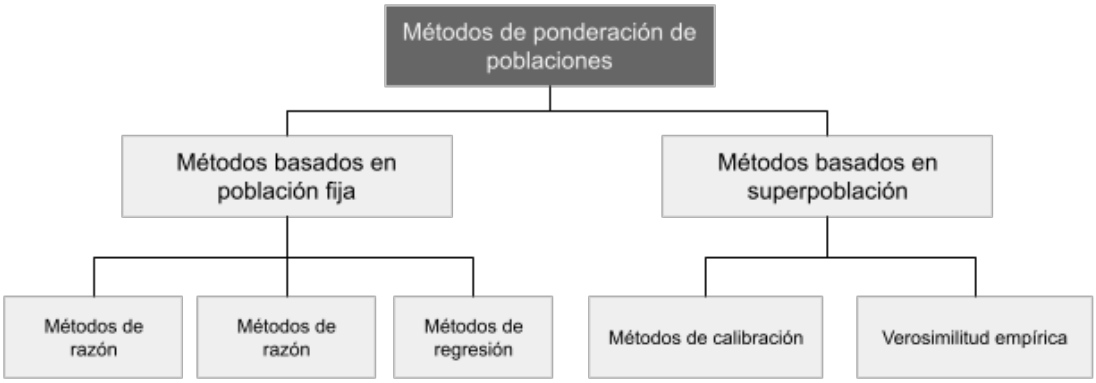
Lohr (2019) propone cuatro mecanismos para controlar la no respuesta: 1) prevención a través de un diseño muestral adecuado; 2) tomar una muestra representativa de los que no responden y usar esta submuestra para realizar una inferencia sobre el resto de los ausentes; 3) usar un modelo para predecir los casos del conjunto que no tiene respuesta; y 4) ignorarla, cuestión que se desaconseja totalmente .

Existen distintos enfoques para realizar la estimación de las ponderaciones poblacionales, resumidos en la Figura 3.6. Los primeros métodos propuestos sobre la base de una población fija, son los de estimación indirecta, entre los que destacan los métodos de razón, de diferencia y de regresión (Fuller, 2011). Todos ello se basan en modificar la forma original del estimador a través de la incorporación de las variables objetivo y auxiliares. Como estos métodos no se garantiza, de forma general, la reducción del error. Para solucionarlo, se

[17] Peso que indica la proporción de registros que representa una observación de la muestra.

propusieron posteriormente los modelos de superpoblación (Pérez-Villalta, 2002; Sánchez-Crespo, 2002) que generalizan los métodos de población fija.

Figura 3.6. Métodos de ponderación de poblaciones



Fuente: elaboración propia.

De forma más concreta, el método de muestreo en poblaciones finitas considera que los valores x_i, de la característica de interés asociados a una unidad u_i de una población finita U son fijos aunque desconocidos (excepto para los elementos de la muestra una vez que ha sido obtenida). Por tanto, esos valores no tienen la consideración de aleatorios. Dicha aleatoriedad, procede completamente de la selección de la muestra y se reflejan en el diseño muestral probabilístico. Los enfoques de población fija se basan en que los estimadores introducen variables indicadoras de la pertenencia de una unidad de la muestra, cuya distribución solo depende del sistema de selección usado.

Cassel (1977) indica que la aleatoriedad observada en una muestra puede tener de tres orígenes distintos:

- El método de selección de las unidades, que es el que se considera en el enfoque de clásico de las investigaciones por muestreo.
- Los métodos de medición de las variables en las unidades seleccionadas.
- El proceso que genera la verdadera medida de la variable para cada unidad, es decir que la fuente tiene cierta estructura aleatoria.

Los modelos de superpoblación, por otra parte, se apoyan principalmente en el uso de variables auxiliares X, cuyos valores son conocidos para todos los individuos de la población Y. Un dato x es un número real que procede de una población tras ser investigada o sujeta a experimentación. Si se repite la investigación o experimento, en general, se obtendría otro dato x', sobre el que si repetimos de en sucesivamente dará lugar a nuevos valores: x'', x''', y así sucesivamente. Estos valores pueden considerarse realizaciones muestrales de cierta variable aleatoria X. Este principio se denomina como muestreo repetido (Azzalini, 2017).

En consecuencia, si un dato es desconocido puede ser considerado una variable aleatoria, y cuando se conozca será una realización de ella. En este enfoque, el tipo de modelo más común es el de los estimadores de regresión, definidos por la expresión:

$$y_k = \mu(X_k) + \varepsilon_k \qquad [3.3]$$

donde ε_k es un error aleatorio, X_k la variable aleatoria e y_k la variable de interés, para la observación k, y μ la función de regresión que relaciona los predictores con la variable de interés.

La perspectiva de ponderación basada en modelos, propone que los estimadores de regresión generalizada son óptimos mientras que la población provenga de una superpoblación que siga un modelo de regresión lineal. Existen dos clases de estimadores basados en modelos: los estimadores de calibración (Deville y Särndal, 1992) y los de verosimilitud empírica (Chen y Qin, 1993; Chen y Sitter, 1999). Estos últimos, proponen un enfoque basado en la probabilidad empírica para el uso de información auxiliar. En nuestro caso, se usará un método de superpoblación basado en modelos a través de un estimador de calibración, que permitirá ajustar los pesos muestrales.

3.2.3.1 Calibración de los elevadores de oferta

La calibración es un método particular de correspondencia estadística que se aplica en estudios basados en encuestas, y que mejora la precisión de la estimación de parámetros sobre la muestra a través de información auxiliar. Es una técnica de uso habitual por las oficinas estadísticas nacionales europeas y norteamericanas.

La técnica se fundamenta en los procedimientos introducidos por Deming y Stephan (1940), y se propone formalmente por Deville y Särdnal (1992). En este trabajo se demuestra la equivalencia asintótica de la calibración[18] para el estimador de regresión generalizada (Cassel *et al.*, 1976), lo que garantiza las propiedades estadísticas de los estimadores calibrados. Posteriormente, Särndal (2007) propone una definición completa del método de calibración, que consiste en el cálculo de los pesos bajo unas restricciones, el cómputo de estimadores lineales ponderados sobre parámetros de la muestra y la construcción de un estimador casi insesgado (eliminando los sesgos de no respuesta y sobre e inframuestreo). Desde entonces su uso ha ido en aumento, y se han desarrollado nuevos métodos consecuencia de disponer computadores con mayor capacidad de cálculo. Entre ellos, los métodos basados en técnicas no paramétricas generales

[18] Para más información sobre el método de calibración, véase el Anexo 3a de este capítulo.

(Wu y Sitter, 2001); las basadas en regresiones locales polinómicas y las redes neuronales (Montanari y Ranalli, 2005); y optimizaciones en función de la medida de distancia[19] (Devaud y Tillé, 2019).

En la presente metodología, la calibración está orientado a ajustar las unidades de oferta y hacerlas corresponder con las unidades de mercado alquiler a lo largo del tiempo. Este proceso persigue la eliminación de sesgos de representación, de forma que todos los estratos de la muestra cumplan el criterio de suficiencia estadística de Fisher (Kullback, 2012). En su estado original, los registros de oferta tienen una serie de desequilibrios importantes, que son entre otros:

- Aún cuando el colectivo de oferta fuera igual al del mercado, el portal Idealista no representa al colectivo total de la oferta, a pesar de su amplia penetración de mercado; o bien que ciertos estratos socioeconómicos menos favorecidos estén menos representados en los portales inmobiliarios (Boeing y Waddell, 2017). Estas cuotas de mercado también varían a lo largo del tiempo (infrarrepresentación).
- El portal no cuenta con ciertos segmentos del alquiler, como alquiler social (falta de respuesta).
- En ciertas zonas, es habitual que el mismo inmueble esté anunciado por varias inmobiliarias (sobrerrepresentación) (Pangallo y Loberto, 2018; Wang *et al.*, 2020); o bien, los inmuebles que suscitan menos interés están sobrerrepresentados en la muestra de oferta (Han y Strange, 2016).

El dato de oferta cuenta con una extracción mensual de los inmuebles en el portal, mientras que las fuentes utilizadas para calibrar los pesos tiene frecuencia anual. Para utilizar la misma escala se agrupan las observaciones por código de anuncio[20] y año.

Las variables comunes X que se utilizan para establecer la correspondencia, deben existir tanto en el fichero Idealista como en la fuente estadística alternativa, y además es deseable que presenten una correlación lo más fuerte posible con las variables de interés Y. Para el caso del índice de precios, las magnitudes de interés serán los precios de oferta y de mercado.

Se realiza un proceso de reponderación en dos etapas (Lohr, 2019), que comienza con una primera calibración de la oferta sobre los totales del censo y una segunda calibración utilizando la información de la EPF. Esta segunda calibración se utiliza para hacer los ajustes temporales de los datos censales, por tanto solo aplica a los años 2012 y siguientes.

[19] Véase Anexo 3a de este capítulo.
[20] Variable Idealista *ADID* del fichero idealista.

Debido a que el nivel de profundidad de la información del que se dispone en el censo es distinto en la ciudad de Madrid que en el resto de municipios, se divide el proceso en dos calibraciones, uno para la ciudad y otro para el resto. Se describe el proceso en la Figura 3.7.

Figura 3.7. Métodos de ponderación de poblaciones

Fuente: elaboración propia.

Durante la construcción de los modelos de calibración se observan problemas de convergencia a partir de 2016, que se pueden la inestabilidad temporal del dato original de la EPF. Aunque no se puede determinar con exactitud la causa de la inestabilidad, ésta podría deberse del cambio metodológico en la EPF introducido a partir del año 2016 (INE, 2006a).

Para evitar los problemas de cambios abruptos en los pesos de la encuesta, se realiza un proceso de ajuste de los pesos poblacionales con suavizado exponencial sobre los totales de la EPF. El ajuste se realiza en función de su relación con los totales del Censo, de manera que, se intenta preservar el dato original pero asegurando una variación limitada con respecto al dato de referencia. La diferencia permitida entre ambas magnitudes es directamente proporcional a la diferencia en años entre 2011 y el año de la EPF que corresponda. De esta forma, se permiten variaciones mayores en años más lejanos al año base que en los años más cercanos.

El método de calibración seleccionado es una regresión logística con bandas[21] (logit). El estimador general de regresión, cuyo acrónimo es *GREG* (Deville y Särndal, 1992), se puede definir de forma coherente en muchas formas diferentes, por ejemplo: lineal, *raking*, truncado y *logit*. Se decide usar el logístico, o exponencial generalizado, porque permite controlar la aparición de valores extremos, es asintóticamente consistente, y ofrece siempre pesos positivos (Folsom y Singh, 2000).

El estimador de calibración generalizado calcula los pesos g, denominados *g-weights*, y calculados como $g_k = F(\lambda' z_k)$, donde z_k es un vector con valores definidos para registro $k \in s$ (o $k \in r$ donde r es el conjunto con respuesta

[21] Se ha utilizado el paquete *sampling* de R (Tillé y Matei, 2016).

conocida) y comparten la dimensión de un vector de variables auxiliares x_k. Los vectores z_k y x_k deben estar fuertemente correlados y el vector λ se determina a través de la ecuación de calibración:

$$\sum_{k \in s} d_k \cdot g_k \cdot x_k = \sum_{k \in U} x_k \qquad [3.4]$$

Si los vectores X_s y Z_s son iguales, se finaliza satisfactoriamente el proceso de calibración. En el caso del método *logit* los valores g, y por tanto los elevadores finales, estarán limitados entre un valor mínimo y máximo (Tillé y Matei, 2016). Los pesos g se pueden expresar también en función de la relación entre los pesos poblacionales finales y la distancia:

$$g_k = \left(\frac{w_k}{d_k} \right) \qquad [3.5]$$

donde w_k y d_k son el peso y la distancia para la observación k.

Dado que no existe un proceso para estimar los límites óptimos *a priori*, se desarrolla un proceso que se inicia con una banda superior e inferior amplias, y que va reduciendo progresivamente con un factor δ variable, por la parte superior e inferior, siguiensdo lo propuesto por Rao (1996). El proceso termina cuando no es posible reducir la parte superior o ampliar la parte inferior del intervalo.

Se utilizan varios niveles de precisión para encontrar el intervalo óptimo, que se muestran en la Tabla 3.2. El proceso comienza reduciendo el intervalo con una precisión baja hasta que es imposible reducir el valor. En ese caso, sobre el último valor válido, se intenta reducir el intervalo de forma sucesiva, hasta no poder mejorar el último intervalo válido.

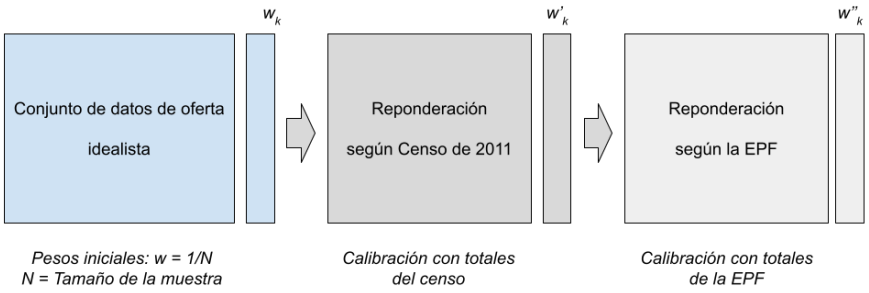
Tabla 3.2. Incrementos máximos y mínimos en bandas de calibración

Precisión	**Delta banda inferior**	**Delta banda superior**
Alta	0,002	-0,100
Media	0,005	-0,500
Baja	0,010	-1,000

Fuente: elaboración propia

En cada uno de los pasos del proceso se estimará un conjunto de pesos para todos los registros de la muestra de oferta, tal y como indica la Figura 3.8.

Figura 3.8. Cálculo de pesos



Fuente: elaboración propia.

Los pesos originales de la oferta w, se sustancian el elevador muestral w_k para cada registro k en el conjunto de datos. Este peso, se ajusta con la calibración en función de los totales del censo, para dar lugar a unos nuevos elevadores muestrales w'_k. Dado que estos pesos elevan la muestra al Censo de 2011, se aplica una nueva calibración para obtener los pesos definitivos de mercado w''_k, usando los totales de la EPF.

Los totales utilizados la calibración del censo han sido la superficie total útil y número de las viviendas en alquiler por estrato. Mientras que en la calibración de la EPF, se han utilizado el número de hogares en alquiler, a partir del factor de elevación de la muestra.

Las restricciones del modelo de calibración intentan mantener el equilibrio de la distribución de los totales y la estructura original de pesos de la oferta. En este sentido, se recuerda que en la calibración censal se controlan los totales por zona, mientras que en la segunda no, al no estar disponibles.

A modo de ejemplo, en la Tabla 3.3, se muestran los distintos pesos[22] que toman cinco anuncios a lo largo del proceso. Partiendo de los w_k originales de la tabla de anuncio se obtienen los w'_k, que son los pesos una vez reponderados según el censo, y finalmente los w''_k, una vez reponderados mediante la EPF. Para el primer anuncio se parte de un elevador muestral de 0,004%, es decir este anuncio representa ese porcentaje con respecto al total de viviendas, que teniendo en cuenta el censo representa a un 0,014% y vuelto a calibrar con la EPF llega a representar un 0,025% de las viviendas totales del colectivo.

[22] En este ejemplo, los pesos se expresan en porcentaje sobre el total poblacional.

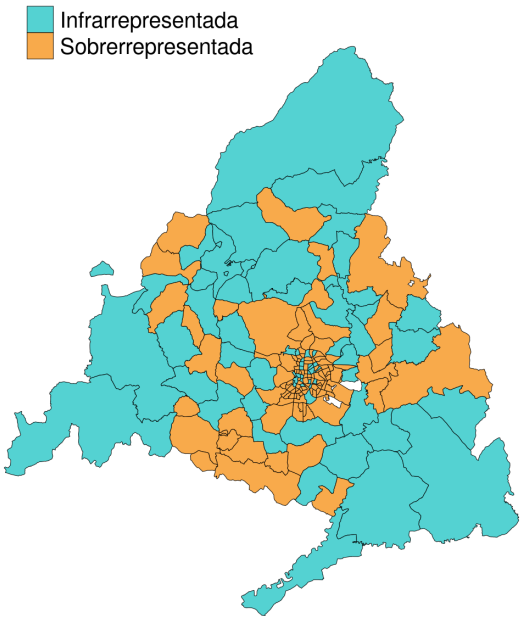
Tabla 3.3. Resultados de la reponderación, cambios en los elevadores muestrales en porcentaje

Código Anuncio	Peso Original	Peso Censo	Peso EPF
321578	0,019	0,069	0,260
2006331	0,019	0,043	0,221
2166585	0,030	0,098	0,198
25122953	0,019	0,044	0,216
25294841	0,023	0,107	0,206

Fuente: elaboración propia

En términos geográficos, la Figura 3.9 representa un mapa sobre el ajuste de los pesos en las distintas zonas para el año 2015. Cada zonas muestra si existe sobrerrepresentación o infrarrepresentación con respecto a su tamaño real. Se observa que principalmente las zonas centrales de la región tienden a la sobrerrepresentación, al contrario de las periféricas. Este fenómeno se podría atribuir al hecho de ser áreas con menor actividad inmobiliaria[23] y por tanto menor rotación de contratos.

Figura 3.9. Zonas sobrerrepresentadas e infrarrepresentadas en oferta



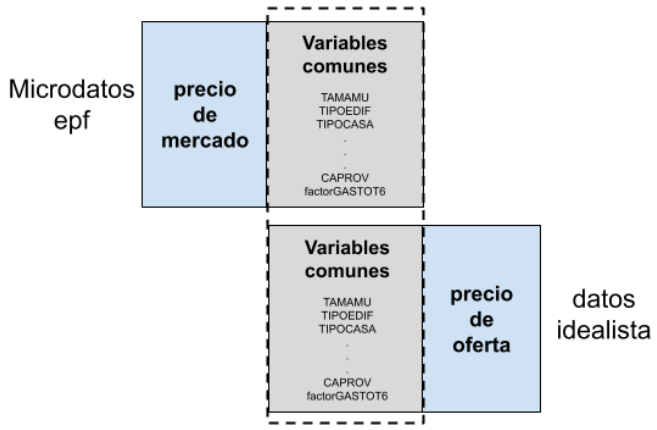
Fuente: elaboración propia.

[23] Según datos de idealista sobre el número de contactos medios por anuncios por zona.

3.3 Modelo de correspondencia de precios

El segundo método de correspondencia estadística es el de los precios, y relaciona los mismos en las bases de microdatos de Idealista y la EPF. Se construye mediante un modelo hedónico que tiene como covariables las características de la vivienda y el precio de oferta, y como variable objetivo, el precio del alquiler.

Figura 3.10. Modelo de correspondencia de precios



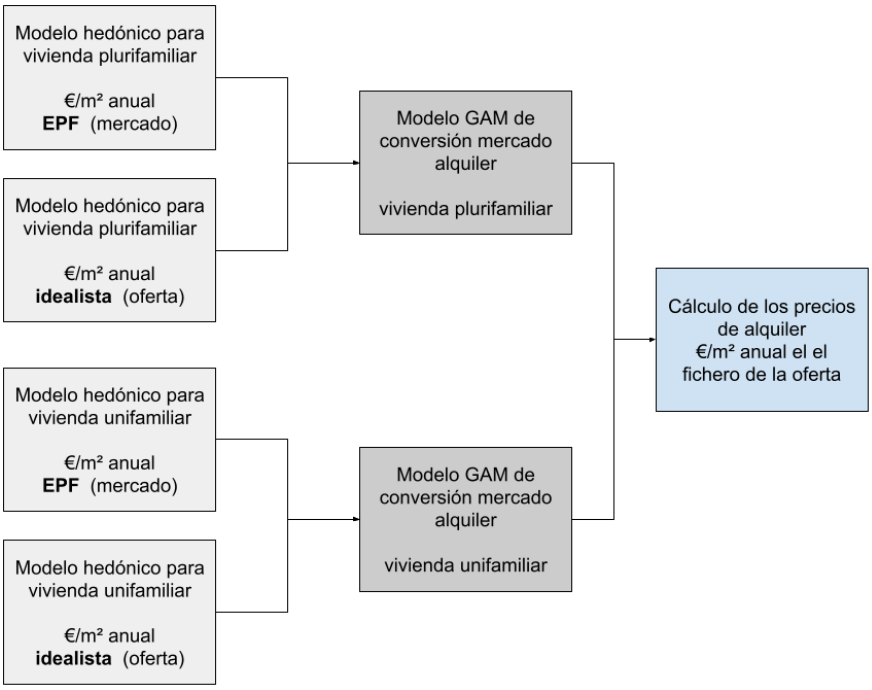
Fuente: elaboración propia.

Al desconocerse el precio de mercado de las viviendas a nivel individual para la oferta, primeramente, se realiza un proceso de imputación de precios de alquiler para todos los registros. Sobre estos datos, se construye un modelo lineal que traduce los precios individuales de oferta a su correspondiente precio de mercado. Este proceso de correspondencia estadística basada en modelos, se apoya en las variables comunes entre los datos de Idealista y la EPF, como se muestra en la Figura 3.10.

El modelo hedónico desarrollado se forma a partir de un conglomerado de 54 modelos (3 modelos x 2 tipos de vivienda x 9 años). El resultado final estima el precio anual del alquiler anual para cada registro del colectivo de estudio, que estará representado por el fichero Idealista con las ponderaciones de mercado.

El proceso completo se resume en la Figura 3.11, el cual se inicia con la creación de dos modelos de imputación de precios de mercado y oferta, que se unen con un último modelo que denominaremos de correspondencia de precios.

Figura 3.11. Descripción proceso de construcción de modelos hedónicos del mercado



Fuente: elaboración propia.

Siguiendo la recomendación de Eurostat (2014), se han desarrollado modelos hedónicos diferentes para cada tipología de vivienda residencial, unifamiliar y plurifamiliar. Ya que las diferencias en precios y características de ambos tipos son notables.

Para los 27 modelos hedónicos iniciales (oferta y alquiler), se han empleado modelos lineales aditivos generalizados[24] (en adelante GAM) (Hastie y Tibshirani, 2017), y árboles de regresión de tipo *Random Forests*[25] (Breiman, 2001). Se opta por los árboles por su mejor adaptación ante las características de la fuente de datos, que son son: la presencia de no linealidades, heterogeneidad espacial y heterocedasticidad (Baldominos *et al.*, 2018; Hastie *et al.*, 2017; Hjort *et al.*, 2022; Hong *et al.*, 2020; Valier, 2020). El Anexo 3c describe con detalle la técnica GAM, mientras que, el Anexo 3b detalla el método *Random Forests*.

El resumen de las técnicas aplicadas para cada hedónico se resume en la Tabla 3.4.

[24] Se ha utilizado el paquete *mgc"* de (Simon, 2017).
[25] Mediante el paquete *"ranger"* de R (Wright y Ziegler, 2015).

Tabla 3.4. Tipos de modelos creados

		Modelos		
Tipo	**Año**	**EPF**	**Idealista**	**Correspondencia**
Plurifamiliar	2011	R. Forests	R. Forests	GAM
	2012	R. Forests	R. Forests	GAM
	2013	R. Forests	R. Forests	GAM
	2014	R. Forests	R. Forests	GAM
	2015	R. Forests	R. Forests	GAM
	2016	R. Forests	R. Forests	GAM
	2017	R. Forests	R. Forests	GAM
	2018	R. Forests	R. Forests	GAM
	2019	R. Forests	R. Forests	GAM
Unifamiliar	2011	GAM	GAM	GAM
	2012	GAM	GAM	GAM
	2013	GAM	GAM	GAM
	2014	GAM	GAM	GAM
	2015	GAM	GAM	GAM
	2016	GAM	GAM	GAM
	2017	GAM	GAM	GAM
	2018	GAM	GAM	GAM
	2019	GAM	GAM	GAM

Fuente: elaboración propia

Para decidir qué método se utiliza en cada caso se ha aplicando el principio de parsimonia, de forma que, a igualdad de condiciones se prefieren lo modelos lineales. Sin embargo, en las viviendas plurifamiliares ha sido necesario utilizar *Random Forests* para lograr un correcto nivel de ajuste. Los motivos se pueden fundamentar en el hecho de que las viviendas unifamiliares representan el grupo más heterogéneo, en características y ámbitos geográficos, de la comunidad de Madrid.

El método *Random Forests* requiere el establecimiento de una serie de hiperparámetros, sobre las características de los árboles utilizados, las variables utilizadas en el proceso de construcción del modelo, el método para estimar la importancia de las variables y los pesos utilizados. Los pesos ponderan la importancia de las observaciones en la muestra, de forma que el método minimiza el error final del modelo ponderado según el peso poblacional de cada una de las

instancias. Los hiperparámetros aplicados se resumen en la Tabla 3.5.

Tabla 3.5. Hiperparámetros del modelos de mercado de tipo Random Forests

Modelo	Tipo	Número de árboles	Mtry	Tamaño mínimo nodo	Tipo importancia	Pesos
Mercado (EPF)	Plurifamiliar	1000	9	12	impurity	EPF

Fuente: elaboración propia

donde el número de árboles indica el número de estimadores que utiliza el algoritmo para calcular la estimación final; el tamaño mínimo de nodo es el número de observaciones mínimas asociadas a cada nodo hoja de los árboles; y el parámetro $mtry$ se refiere al número de variables sobre las que se puede aplicar una regla de decisión al construir los árboles. El tipo de importancia recoge al criterio de reducción de entropía[26] utilizado para establecer los cortes.

3.3.1 Modelos hedónicos básicos de mercado y oferta

Por cada tipo de vivienda se desarrolla un par de modelos hedónicos sobre un conjunto de variables comunes de las fuentes, Idealista y EPF. En el primer caso, se estima el precio del alquiler a precios de mercado, utilizando el conjunto de microdatos de la EPF[27]. La variable a predecir es el logaritmo del precio del alquiler anual por metro cuadrado útil, siendo su forma funcional la siguiente:

$$\begin{aligned} log(\hat{P}_m) \leftarrow & TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + \\ & SUPERF + ANNOCON + DENSI + INTERINPSP + \\ & NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.6]$$

donde $\hat{P}_m$ es el precio de alquiler anual de mercado por unidad de superficie útil, $TAMAMU$ el tamaño del municipio, $TIPOEDIF$ el tipo de edificio, $TIPOCASA$ el tipo de vivienda, $ZONARES$ el tipo de zona residencial, $SUPERF$ la superficie útil de la vivienda en m², $ANNOCON$ el año de construcción, $DENSI$ la densidad de población de la zona, $INTERINPSP$ los ingresos del cabeza de familia, $CCAA$ la comunidad autónoma, $factorGASTOT6$ los ingresos del cabeza de familia, $CAPROV$ es una variable dicotómica que indica si la observación está en la capital de provincia o no.

[26] En este caso la entropía se interpreta como el nivel de desorden en la variable objetivo, medido como índice Gini, que reduce el árbol al dividir la población con esta variable.
[27] Ver sección 2.4.1.

Posteriormente, se construye un modelo que estima la misma magnitud pero para la oferta ($\hat{P}_o$), en cuyo caso la fuente es el fichero Idealista sobre las mismas variables del modelo anterior, como se aprecia en su forma funcional:

$$\begin{aligned} log(\hat{P}_o) \leftarrow TAMAMU + TIPOEDIF + TIPOCASA + ZONARES + \\ SUPERF + ANNOCON + DENSI + INTERINPSP + \\ NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.7]$$

3.3.2 Modelos de correspondencia de precios

Los dos conjuntos de precios, de mercado ($\hat{P}_m$) y oferta ($\hat{P}_o$), deben relacionarse a través de un modelo de correspondencia. El enlace se realiza también mediante un modelo, cuyo conjunto de variables comunes son las covariables presentes en ambos modelos más el precio de oferta. Al existir relaciones no lineales entre el precio de mercado y las covariables, se ha decidido utilizar un modelo lineal GAM, que ofrece un buen balance entre interpretabilidad y ajuste (Hastie y Tibshirani, 2017), y evita algunos inconvenientes de los modelos de aprendizaje estadístico aplicados al precio de la vivienda (Nguyen y Cripps, 2001), como el la selección correcta de hiperparámetros.

Las funciones de suavizado del modelo GAM permiten adaptar las contribuciones de las covariables a sus diferentes valores. El modelo de correspondencia de precios se especifica de la forma siguiente:

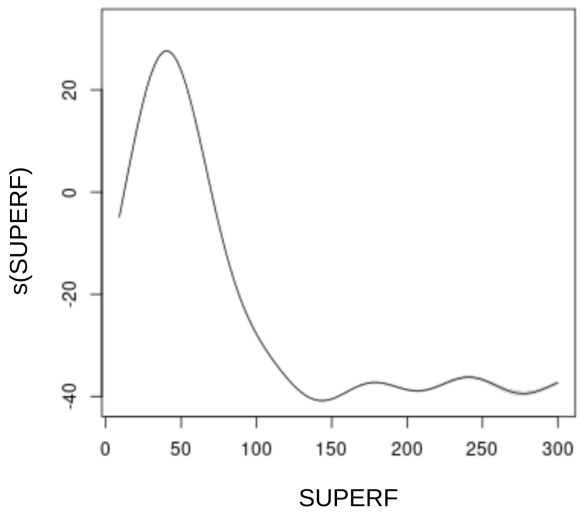
$$\begin{aligned} log(\hat{P}_m) \leftarrow s(\hat{P}_o) + TAMAMU + TIPOEDIF + TIPOCASA + \\ ZONARES + s(SUPERF) + ANNOCON + DENSI + INTERINPSP + \\ NHABIT + CCAA + CAPROV + factorGASTOT6 \end{aligned} \quad [3.8]$$

donde el término $s(SUPERF)$ indica que se aplica una función de suavizado, o *spline*, para modelar la relación entre la superficie y el precio. Las funciones de suavizado s se calculan mediante la agregación de una serie de funciones base (Hastie *et al.*, 2017), y son una generalización de los modelos lineales. De forma que una regresión por mínimos cuadrados usa una función de suavizado s en lugar de un coeficiente fijo, en cuyo último caso representarse geométricamente como una recta con pendiente constante (Hastie y Tibshirani, 2017) que representa la contribución del parámetro, en lugar de una curva en el caso de la mencionada s.

La Figura 3.12 muestra los valores de la suavizado s para el la variable *SUPERF*

en el modelo de correspondencia oferta-mercado. La función se puede interpretar como un coeficiente dinámico para los distintos valores de la variable. En la Figura se observa como para superficies pequeñas el coeficiente es positivo hasta llegar a 50 m² aproximadamente. Entre 50 m² y 80 m², la contribución es positiva pero decreciente, mientras que a partir de 80 m², la contribución al precio es negativa. Para valores superiores la relación es decreciente, hasta que los 140 m², donde los incrementos en la superficie no implican incrementos significativos en el precio.

Figura 3.12. Valores de la función de suavizado (s) para la variable superficie útil en modelo de correspondencia



Fuente: elaboración propia.

El modelo de correspondencia permite estimar el precio de mercado para cualquier registro de la oferta. Por tanto, dado el alto nivel de detalle de los atributos del conjunto de datos de la oferta y los elevadores muestrales, es posible construir series de precios de mercado altamente desagregadas. Sin embargo, el conjunto de atributos utilizados para estimar los precios de oferta no es muy exhaustivo, por lo que el modelo adolece sesgos por omisión de variable. Para solucionar esta cuestión y lograr un mayor nivel de precisión en las series de precios, será necesario ajustar el resultado del modelo de correspondencia con un modelo hedónico más preciso (que es el hedónico de oferta que se presentará en el Capítulo 5).

Por otra parte, el método de correspondencia no utiliza como variables de áreas geográficas, ya que los microdatos de la EPF no disponen de dicha información. Debido a que las cada zona tiene un comportamiento particular, será necesario un proceso posterior de control para asegurar el correcto control de la heterogeneidad espacial sobre los resultados.

3.4 Resultados

El proceso de correspondencia asigna, para cada registro de la oferta, un elevador muestral y un precio de mercado. Para evaluar la calidad de los resultados se estudiará el cumplimiento de las condiciones propuestas por Rässler (2012):

- Preservar los valores individuales: se comprueba que se respetan las sumas de los pesos totales para los distintos criterios de estratificación, además de asegurar que las diferencias entre el peso original y el final se encuentran dentro de bandas aceptables.
- Preservar la estructura de la correlación: se analiza si el precio de mercado estimado mantiene un buen nivel de ajuste con respecto a sus valores originales. Para ello, se estudiará la calidad del modelo de correspondencia de precios.
- Preservar la distribución conjunta: se valida si la distribución conjunta de los pesos muestrales finales no difiere, de forma sensible, con respecto a la distribución conocida del fichero de la EPF.
- Preservar las distribuciones marginales: analiza que las distribuciones marginales de variable objetivo se corresponden con las originales. De forma particular, se estudia el efecto en la distribución espacial del modelo, dado que, el proceso de reponderación utiliza parcialmente esta información para la estimación de los pesos.

3.4.1 Valores individuales

Los pesos individuales de la calibración deben cumplir los requisitos de ajuste, es decir, ser positivos y estar acotados entre un valor máximo y mínimo. Al cumplirse la condición de convergencia en el proceso de calibración, expresada como el máximo valor de la diferencia entre los totales, se asegura que la suma de los totales de cada una de las variables no presentan diferencias apreciables. El criterio de convergencia asegura la siguiente condición:

$$\frac{max(X_s \cdot g_s \cdot d_s - T_s)}{T_s} \leq 10^{-6},\ \forall s \in S \qquad [3.9]$$

donde T_s representa el total para cualquier estrato s de los S criterios de estratificación; X_s son los valores individuales de las variables; y $g_s \cdot d_s$ los nuevos pesos muestrales. La condición anterior asegura que el ratio entre la diferencia de los totales ponderados ($X_s \cdot g_s \cdot d_s$) con respecto a los totales T_s es mayor que uno entre un millón.

La condición anterior se cumple para los límites máximo y mínimo de las g

distancias en cada uno de los procesos de calibración, calculados de forma iterativa (Rao, 1996). La Tabla 3.6 contiene las bandas finales de calibración[28] para el Censo, siendo el rango más amplio de 0,05, para la parte inferior, a 10 para la superior. Se observa que el rango se amplia, especialmente para los años 2018 y 2019, pero incluso en este caso no representan *g-distancias* extremas (D'Orazio *et al.*, 2006). En términos zonales, no se aprecia diferencias importantes entre Madrid y el resto de la provincia.

El factor de 10 en 2019, significa que un registro de oferta representaría 10 hogares en alquiler. Para un factor de 0,05, en la banda inferior, se necesitarían 20 registros de oferta para representar un registro del mercado.

Tabla 3.6. Bandas de calibración Censo y EPF

	Censo				EPF	
	Madrid		Resto CAM		Todas las zonas	
Año	**Inferior**	**Superior**	**Inferior**	**Superior**	**Inferior**	**Superior**
2011	0,28	4,50	0,27	6,67		
2012	0,24	6,00	0,24	6,95	0,46	1,90
2013	0,20	5,00	0,20	5,00	0,58	2,00
2014	0,25	4,00	0,20	3,00	0,28	3,00
2015	0,26	4,00	0,16	4,50	0,38	4,00
2016	0,18	3,00	0,18	3,25	0,38	3,00
2017	0,10	3,50	0,10	4,50	0,08	5,00
2018	0,05	10,00	0,05	10,00	0,10	5,50
2019	0,05	10,00	0,05	10,00	0,10	5,50

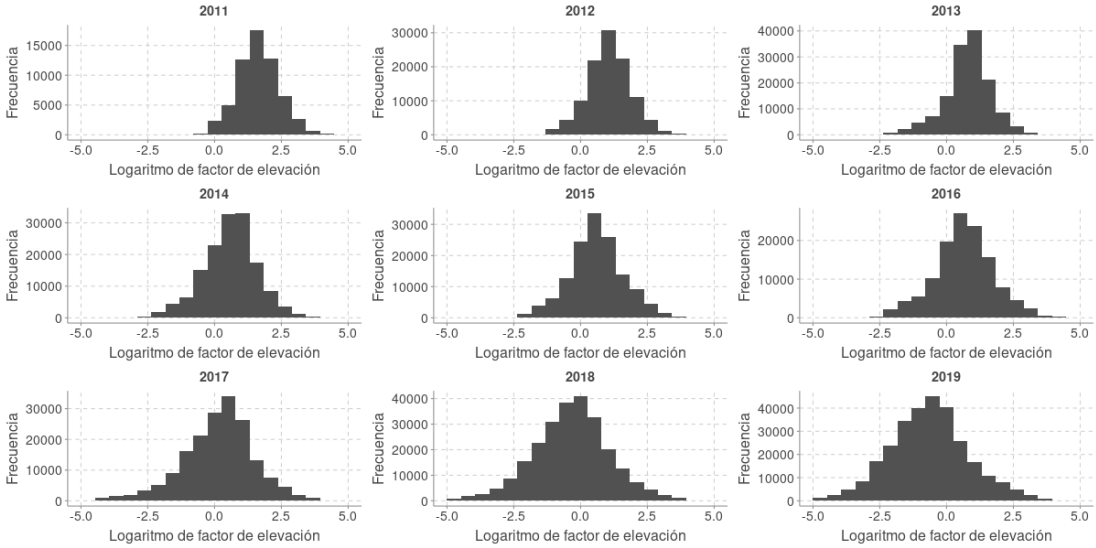
Fuente: elaboración propia

En la calibración de la EPF las bandas son más estrechas que en el caso del Censo. Sin embargo, como sucede en el caso anterior, los intervalos se van ampliando en los últimos años de la serie, lo cual es lógico al estar más alejados del periodo base.

Como se ha indicado anteriormente, la EPF muestra cierta inconsistencia temporal a partir del 2016 que no se corresponden a cambios en el mercado, y que por tanto se puede atribuir a las modificaciones metodológicas introducidas a partir de ese año. En el epígrafe del Anexo 3d, se adjuntan la Tabla 3.17 de totales originales, y la Tabla de 3.18 totales suavizados , en ellas confirma un notable cambio de escala en el año 2016. Los datos una vez ajustados reducen esta discontinuidad, por ejemplo, la variación del total poblacional original entre 2015 y 2016 es del 12%, mientras que el suavizado la reduce al 2%.

[28] Valores mínimos y máximos del factor $g \cdot d$ que ajusta los pesos originales.

Figura 3.13. Distribución de los factores de elevación en escala logarítmica



Fuente: elaboración propia.

En cuanto a la distribución de los pesos muestrales a lo largo del tiempo, se observa que el rango de los factores de elevación finales tiende a ampliarse en los últimos años de la serie (véase Figura 3.13). Además, se observa como los elevadores inferiores a 1 son prácticamente inexistentes en 2011, mientras que en 2019 son mayoritarios, representando más de la mitad de los casos.

Los cambios en la forma de la distribución pueden tener diversos orígenes. Uno específico por la composición de los anuncios publicados en el portal, que de forma progresiva va ampliando su penetración en el mercado. Por otra parte, al tratarse de un mercado en expansión, la relación entre oferta y mercado varía en el tiempo (Ardila *et al.*, 2021; De Wit *et al.*, 2013; Han y Strange, 2016). Ambos motivos dan lugar a que la relación *stock* en alquiler / *stock* en oferta sea progresivamente menor, produciendo elevadores muestrales decrecientes, tal y como vemos en la Figura 3.13, donde el centro de masa de las distribuciones pasa de valores positivos en 2011 a valores ligeramente negativos en 2019.

3.4.1.1 Implicaciones de la estabilidad temporal de la muestra

Con el objetivo de comprobar que la estratificación de origen es comparable a lo largo del tiempo[29], se ha realizado un análisis preliminar para evaluar si la estructura de estratos se mantiene estable a lo largo del tiempo. Para ello, se han tomado las poblaciones de Tres Cantos y Fuentelsaz del Jarama, por representar realidades inmobiliarias distintas. La primera, es una población residencial de la zona metropolitana con un nivel de ingresos medio-alto; y la segunda, una zona

[29] Debido a los cambios metodológicos aplicados sobre la EPF en el periodo de análisis.

rural. En la Tabla 3.7 y la Tabla 3.8 se muestra la presencia de los estratos a lo largo del tiempo (con una “X” si está presente en la muestra y vacío cuando está ausente). En ambos municipios se observan cambios de composición a partir del año 2015 y 2016, que, como se comentaba anteriormente, es el momento en el que la EPF aplica una nueva metodología de trabajo.

Para el caso de Fuentelsaz, solo se dispone de información para el segmento de densidad de población intermedia hasta 2013, mientras que, la zona diseminada mantiene una estructura estable a lo largo del tiempo. Este comportamiento puede ser atribuible a que su población es rural y el tipo diseminado es el mayoritario, por tanto, existe soporte de datos en todos los periodos.

En Tres Cantos, sin embargo, se produce un cambio en la estructura de los estratos que representan el municipio entre 2016 y 2017. Antes de 2017, la muestra se encuentra en zonas densamente pobladas del municipio, mientras posteriormente pasa a concentrarse en zonas con densidad intermedia. Esto en términos intramunicipales no debería tener un efecto importante, pero si lo podría tener en el cálculo de los totales agregados, por densidad de población, en la Comunidad de Madrid.

Este análisis se ha realizado únicamente en los municipios de la Comunidad de Madrid, a excepción de la capital, por cuanto la muestra en la última se mantiene estable.

Tabla 3.7. Presencia de estratos en Fuentelsaz del Jarama (todos excepto casa económica)

Densidad	Tipo	Edificio	Hab.	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
diseminada	Casa media	10 ó más	1 o 2	Urbana media									X
diseminada	Casa media	menos de 10	1 o 2	Urbana alta						X			
diseminada	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Casa media	10 ó más	3	Urbana media			X	X	X	X	X	X	
diseminada	Casa media	menos de 10	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Casa media	menos de 10	4	Urbana media		X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	1 o 2	Urbana media					X	X	X	X	
diseminada	Chalé o casa grande	independiente	1 o 2	Urbana alta						X			
diseminada	Chalé o casa grande	independiente	1 o 2	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	3	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	4	Urbana alta						X			
diseminada	Chalé o casa grande	adosada	4	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	4	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	adosada	5 o más	Urbana alta						X			
diseminada	Chalé o casa grande	adosada	5 o más	Urbana media	X	X	X	X	X	X	X	X	X
diseminada	Chalé o casa grande	independiente	5 o más	Urbana media			X					X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X					
intermedia	Casa media	10 ó más	3	Urbana media			X	X					
intermedia	Casa media	menos de 10	3	Urbana media	X	X	X	X					
intermedia	Casa media	menos de 10	4	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	1 o 2	Urbana media	X		X	X					
intermedia	Chalé o casa grande	independiente	1 o 2	Urbana media	X	X							
intermedia	Chalé o casa grande	adosada	3	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	independiente	3	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	4	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	adosada	5 o más	Urbana media	X	X	X	X					
intermedia	Chalé o casa grande	independiente	5 o más	Urbana media				X					

Fuente: elaboración propia

Tabla 3.8. Presencia de estratos, municipio de Tres Cantos (casa de tipo medio)

Densidad	Tipo	Edificio	Hab.	Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
densa	Casa media	10 ó más	1 o 2	Urbana alta	X	X	X	X	X	X			
densa	Casa media	10 ó más	1 o 2	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	1 o 2	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	1 o 2	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	3	Urbana alta	X	X	X	X	X	X			
densa	Casa media	10 ó más	3	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	3	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	3	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	4	Urbana alta		X	X	X	X	X			
densa	Casa media	10 ó más	4	Urbana media	X	X	X	X	X	X			
densa	Casa media	menos de 10	4	Urbana alta	X	X	X	X	X	X			
densa	Casa media	menos de 10	4	Urbana media	X	X	X	X	X	X			
densa	Casa media	10 ó más	5 o más	Urbana media				X	X	X			
densa	Casa media	menos de 10	5 o más	Urbana alta				X					
densa	Casa media	menos de 10	5 o más	Urbana media		X	X						
intermedia	Casa media	10 ó más	1 o 2	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	1 o 2	Urbana media							X	X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana alta							X	X	X
intermedia	Casa media	menos de 10	1 o 2	Urbana media							X	X	X
intermedia	Casa media	10 ó más	3	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	3	Urbana media							X	X	X
intermedia	Casa media	menos de 10	3	Urbana alta							X		X
intermedia	Casa media	menos de 10	3	Urbana media							X	X	X
intermedia	Casa media	10 ó más	4	Urbana alta							X	X	X
intermedia	Casa media	10 ó más	4	Urbana media							X	X	X
intermedia	Casa media	menos de 10	4	Urbana alta							X	X	X
intermedia	Casa media	menos de 10	4	Urbana media							X	X	X
intermedia	Casa media	10 ó más	5 o más	Urbana media							X	X	X
intermedia	Casa media	menos de 10	5 o más	Urbana media									X

Fuente: elaboración propia

3.4.2 Estructura de la correlación

Para evaluar que la estructura de correlación entre la muestra y los resultados del modelo se mantiene constante, se evaluará la calidad del ajuste de los modelos construidos[30], medida en R^2 ajustado. Se toma esta métrica porque es aplicable tanto a modelos lineales como a árboles de regresión, y es más informativa que otras medidas, como el error cuadrático medio o el error medio absoluto (Chicco *et al.*, 2021).

El R^2 es una medida estadística que recoge la proporción de la varianza de la variable dependiente explicada por un modelo sobre las variables independientes. Toma valores entre $-\infty$ y 1, donde 1 es el ajuste perfecto de un modelo capaz de capturar cualquier comportamiento de la variable objetivo. Mientras que 0 indica que el modelo no puede capturar ninguna relación entre los regresores y la variable objetivo, y los valores inferiores a cero muestran un modelo que ajusta peor que una linea horizontal (Chicco *et al.*, 2021).

De forma general, el R^2 se puede calcular según la expresión:

$$R^2 = 1 - \frac{\sigma^2_{error}}{\sigma^2_y} \qquad [3.10]$$

donde σ^2_{error} es la varianza de los errores del modelo, y σ^2_y es la varianza de la variable dependiente y.

En nuestro caso se opta por una versión ajustada del R^2, que soluciona dos problemas importantes de la medida original: el primero, reduce el sobreajuste asociado un número elevado de grados de libertad del modelo; y el segundo, la propensión del R^2 a ofrecer valores más altos cuanto mayor es el número de parámetros del modelo. Intenta, además, que la magnitud exprese el porcentaje de la variable explicado solamente por los regresores que afectan a la variable dependiente. Por tanto, la versión ajustada del coeficiente de determinación sería:

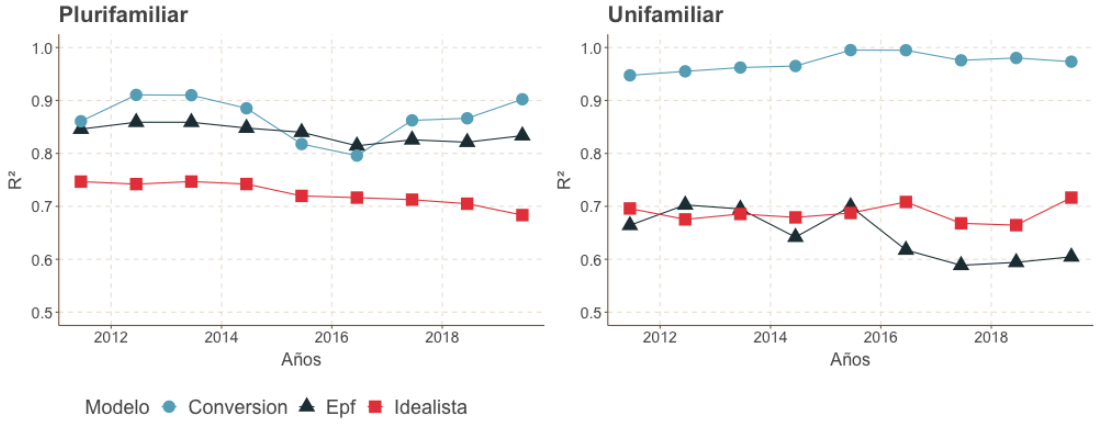
$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \qquad [3.11]$$

donde n son grados de libertad o número de observaciones, y k el número de regresores independientes.

[30] La calidad del ajuste representa la capacidad del modelo a predecir cualquier valor de la distribución de valores de la variable objetivo.

En los modelos lineales, el R^2 se estima de forma directa, mientras que en los de tipo *Random Forests*, se calcula sobre la muestra *out of bag*[31]. Esta última, es un conjunto de observaciones no utilizadas en la construcción de cada árbol[32] .

Figura 3.14. Coeficiente de determinación R²



Fuente: elaboración propia.

La Figura 3.14 muestra los niveles de ajuste de los modelos en términos de R^2, observándose que los niveles de ajuste son muy altos (en general mayores a 0,7), en particular para los modelos de correspondencia en ambos tipos (con una media en torno a 0,9). También se aprecia que en los modelos de EPF e Idealista los valores más bajos se corresponden a las viviendas unifamiliares, en comparación con las plurifamiliares. El menor ajuste de las primeras se puede deber a la ausencia de variables importantes en el modelo, como por ejemplo, el tamaño de la parcela o la limitación máxima de la superficie a 300 m².

En todo caso, un valor del coeficiente de determinación superior al 0,9 es consistente con los resultados de otros autores que utilizan métodos no paramétricos, como por ejemplo, el caso de Ho (2021) para Hong Kong o el Rico y Taltavull (2021) sobre precios de tasaciones en Alicante. Es importante destacar el grado de ajuste de nuestro caso puede considerarse muy bueno, dado que el número de atributos utilizado para la construcción del modelo de mercado es mucho más limitado que el de las publicaciones citadas.

La Tabla 3.9 muestra con mayor detalle los valores de R^2 obtenidos al ajustar los modelos.

[31] Para más detalle sobre el proceso de muestreo, véase el Anexo 3b del presente capítulo.
[32] En el proceso de *bagging* se divide el conjunto de datos en dos partes, *in bag* que se refiere a las instancias usadas para entrenar el modelo, y *out of bag*, en adelante OOB, que se usa para medir el ajuste y error del modelo.

Tabla 3.9. Resumen de ajuste de los modelos de mercado

	Plurifamiliar			Unifamiliar		
Año	**EPF**	**Idealista**	**Corresp.**	**EPF**	**Idealista**	**Corresp.**
2011	0,85	0,75	0,86	0,66	0,70	0,95
2012	0,86	0,74	0,91	0,70	0,68	0,95
2013	0,86	0,75	0,91	0,70	0,69	0,96
2014	0,85	0,74	0,89	0,64	0,68	0,97
2015	0,84	0,72	0,82	0,70	0,69	1,00
2016	0,81	0,72	0,80	0,62	0,71	0,99
2017	0,83	0,71	0,86	0,59	0,67	0,98
2018	0,82	0,70	0,87	0,59	0,66	0,98
2019	0,83	0,68	0,90	0,60	0,72	0,97

Fuente: elaboración propia

Para facilitar la lectura de los resultados de los modelos GAM de correspondencias, se han construido las Tablas 3.11 y 3.10, en las que se representan el signo[33] del coeficiente con su significatividad[34] . En ellas se observa que aquellos coeficientes que son más significativos suelen mantener consistencia de signo a lo largo del tiempo . Además, las dos tipologías muestran comportamientos muy diferentes, las viviendas plurifamiliares cuentan con un mayor grado de significatividad en sus coeficientes y consistencia temporal en los signos de los coeficientes .

Para el caso de las viviendas unifamiliares (Tabla 3.10) la consistencia en términos de signo y significatividad es menor, debido a tener una muestra más pequeña e inestable (véase epígrafe 2.4.3). En términos de mayor estabilidad temporal se pueden destacar los coeficientes de tipo de zona y los de número de habitaciones. Por otra parte, se observan diferencias a lo largo del tiempo, por ejemplo, los años 2015 y 2016 muestran niveles de significatividad mayores que en 2013, 2014 y 2017, en los que los coeficientes no son significativos. Es particularmente interesante destacar que aún así estos tres periodos tienen un R^2 era mayor a 0,9.

Adicionalmente, en el Anexo 3e se adjuntan ejemplos de los coeficientes de los modelos GAM de la EPF y oferta para viviendas unifamiliares. Se observa que el modelo sobre la EPF es mucho más débil en términos de significatividad, siendo las covariables más representativas: el tipo de edificio, tamaño del municipio, número de habitaciones, comunidad autónoma, si está o no en la capital de provincia, y un muy alto nivel de gasto familiar.

[33] "+" expresa que el signo del coeficiente es positivo y "-" que es negativo
[34] Se representan en función de p-valor: *** < 0.001, ** < 0.01, * < 0.05 y "." < 0.1

Tabla 3.10. Signo y significancia de coeficientes modelo GAM correspondencia para viviendas unifamiliares

	Signo									Significatividad								
Coeficiente	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
INTERCEPT	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	-	+	-	+	+	-	+	+	+				***	***		**	***	*
TAMAMUMunicipio con 20.000 o más y menos de 50.000	+	-	-	+	+	+	+	+	+	**			***	**		**	***	*
TAMAMUMunicipio con 10.000 o más y menos de 20.000	-	-	-	+	-	-	-	-	-	***	***	**	***	**	***	**		***
TAMAMUMunicipio con menos de 10.000 habitantes	-	-	-	+	-	+	-	-	-		***	.	***	***		***	***	**
TIPOEDIFVivienda unifamiliar adosada o pareada	-	+	-	+	+	-	+	+	-			**	**		***	**	***	.
ZONARESUrbana alta	+	+	+	-	+	+	+	-	+			**	*		***		***	
ZONARESUrbana media	-	+	+	-	-	+	-	-	-	*			**	***	***		***	.
ZONARESUrbana inferior	-	+	+	-	+	+	-	-	-	*		*	.	***	***		***	**
ANNOCONHace 25 ó más años	-	-	-	-	-	-	-	+	+		*	***	**	***	***	***	***	***
DENSIZona intermedia	-	+	-	-	-	+	-	+	+	***			***		***		***	***
DENSIZona diseminada	-	-	-	-	-	-	-	-	+	***	***	.	***	***	*	***	***	
INTERINPSPDe 500 a menos de 1000 €	+	+	-	+	+	-	+	+	-						***	***	***	
INTERINPSPDe 1000 a menos de 1500 €	+	+	-	+	+	-	+	+	+		*		**	***	***	***	***	
INTERINPSPDe 1500 a menos de 2000 €	+	+	-	+	+	-	+	+	+		.		***	***	***	***	***	
INTERINPSPDe 2000 a menos de 2500 €	-	+	-	+	+	-	+	+	-		*		**	***	***	***	***	
INTERINPSPDe 2500 a menos de 3000 €	-	+	-	+	+	-	+	+	+			**	**	***	***	**	**	*
INTERINPSP3000 o más €	-	+	-	+	+	-	+	+	+			**	*	**	***	**		**
NHABIT3 habitaciones	-	-	-	-	-	-	+	+	+	*	***	***		***	***	**	***	***
NHABIT4 habitaciones	-	-	-	+	-	-	+	+	+		***	***	*	***	***		***	**
NHABIT5 o más habitaciones	-	-	-	+	+	-	+	+	+	***	***	***			***			
CAPROVNo	+	-	+	-	-	-	-	-	-		**		***	***	***	***	***	***
factorGASTOT_1	-	-	-	+	+	-	+	+	-			*	***	***	***	***		
factorGASTOT_2	+	-	+	+	+	-	+	+	+	**	.		***	***	***	***	**	
factorGASTOT_3	+	+	+	+	+	-	+	+	+	**			***	***	***	***	*	
factorGASTOT_4		+	+	+	+						***	.	***	***				

Fuente: elaboración propia

Tabla 3.11. Signo y significancia de coeficientes modelo GAM correspondencia para viviendas plurifamiliares

	Signo									Significatividad								
Coeficiente	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
INTERCEPT	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TAMAMUMunicipio con 10.000 o más y menos de 20.000	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***		***	***
TAMAMUMunicipio con menos de 10.000 habitantes	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TIPOEDIFCon 10 ó más viviendas	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
TIPOCASACasa económica o alojamiento	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
ZONARESUrbana alta	+	-	+	+	-	-	-	-	-	***	***	***		***	***	***	***	***
ZONARESUrbana media	-	-	-	-	-	-	-	-	-		***	*	***	***	***	***	***	***
ZONARESUrbana inferior	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
ANNOCONHace 25 ó más años	+	+	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
DENSIZona intermedia	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
DENSIZona diseminada	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
INTERINPSPDe 500 a menos de 1000 €	-	+	-	+	-	-	-	+	-		***	***	***	***	***	***	***	***
INTERINPSPDe 1000 a menos de 1500 €	+	+	+	+	-	-	-	+	-	***	***	***	***	***	***	***	***	
INTERINPSPDe 1500 a menos de 2000 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	*	***	***
INTERINPSPDe 2000 a menos de 2500 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
INTERINPSPDe 2500 a menos de 3000 €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
INTERINPSP3000 o más €	+	+	+	+	+	+	+	+	+	***	***	***	***	***	***	***	***	***
NHABIT3 habitaciones	-	-	-	-	-	-	-	+	+	***	***	***	***	***	***	***	***	***
NHABIT4 habitaciones	-	+	+	+	-	-	-	-	-	***	***	***	***	***	***	***	***	**
NHABIT5 o más habitaciones	-	+	-	+	-	+	-	-	-	***	***		***	**		***	***	
CAPROVNo	-	-	-	-	-	-	-	-	-	***	***	***	***	***	***	***	***	***
factorGASTOT_1	-	+	+	-	+	-	+	-	+	***	***	***	***	***	*		***	***
factorGASTOT_2	+	+	+	-	+	+	+	-	+	***	***	***	***	***	***		***	***
factorGASTOT_3	+	+	+	+	+	+	+	-	+	***	***	***		***	***		***	***
factorGASTOT_4		+	+	+	+	+	+	-			***	***	***	***	***	*	***	

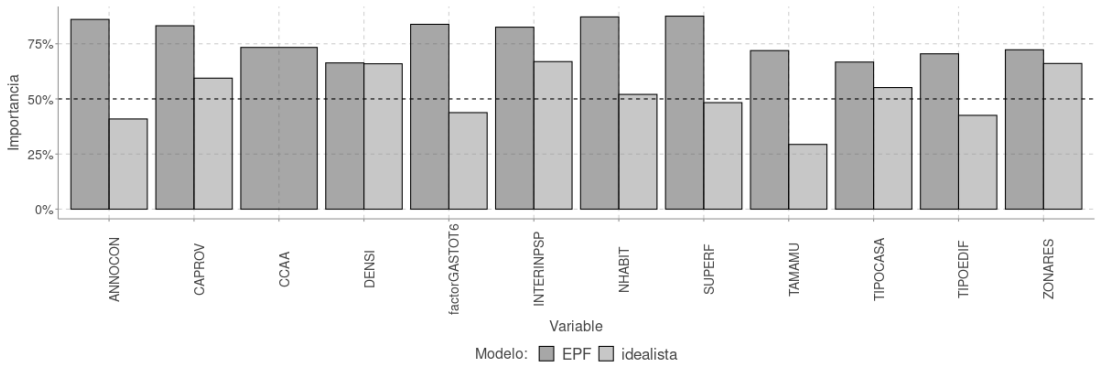
Fuente: elaboración propia

Dado que para los modelos construidos con *Random Forests* no es posible estimar una significatividad de los coeficientes, se puede analizar la contribución de las variables en función de su capacidad de eliminación de la entropía o "impureza"[35]. Esta medida se puede entender como la importancia de un predictor para reducir la variabilidad de los residuos del modelo (que por otra parte es el principio sobre el que se basa el R^2).

Tanto el peso como la significatividad de los coeficientes permiten estudiar el efecto de las variables auxiliares en el proceso de correspondencia en los modelos de tipo no paramétrico. Es conveniente analizar ambas medidas, ya que generalmente su eficacia depende del caso (Zhang y Nguyen, 2020).

Por otra parte, en Figura 3.15 se muestra la importancia de las 12 variables más importantes. Para cada una de las variables del modelo de la EPF se indica su nivel de importancia normalizada con respecto a su aporte en el modelo[36]. Se observa que todos los casos los valores son muy altos, lo que indica que son significativas e intervienen casi en la misma medida en la construcción de los árboles de decisión.

Figura 3.15. Importancia de las variables para los modelos EPF e idealista para viviendas plurifamiliares



Fuente: elaboración propia.

En el modelo de la EPF destacan como variables con más peso: el año de construcción, la provincia, el número de habitantes del municipio, la provincia y los factores de gastos e ingresos. Para el modelo de oferta de Idealista, las variables más importantes son la provincia, densidad de población, nivel de ingresos y tipo de zona residencial, confirmando la importancia de la zona en el modelo hedónico.

A la vista de los resultados de ajuste y de importancia de variables, se puede concluir que la selección de variables auxiliares para la calibración cumple los

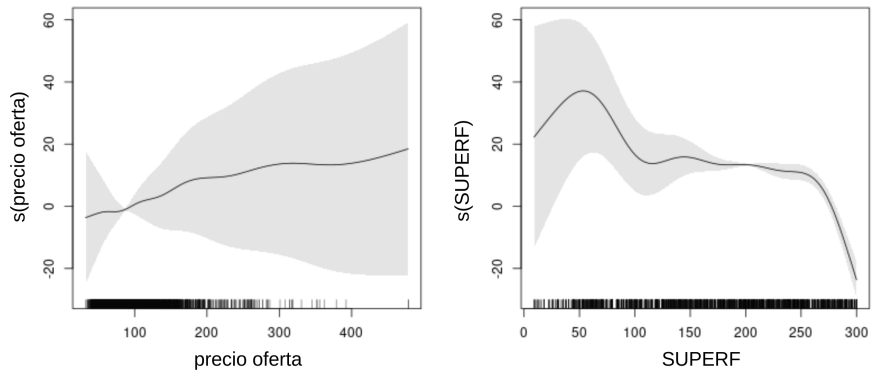
[35] En la configuración utilizada se ha utilizado *impurity* como medida de importancia de variables.

[36] La importancia se calcula se calcula en términos de "impureza" y representa la varianza que reducen los cortes en los que interviene una variable, este valor se normaliza diviendolo por la máxima medida de impureza entre todas las variables.

tres criterios de Särndal y Lundström (2008): 1) explicar la variable respuesta; 2) que las covariables sean todas significativas; y 3), servir para desarrollar la estratificación del índice de precios.

En importante resaltar que los intervalos de confianza de las funciones de suavizado de los modelos GAM, en las viviendas unifamiliares, son muy amplios (representados en la Figura 3.16 el color gris alrededor de la línea de regresión). Lo cual puede relacionarse un menor nivel de grado de R^2. En el caso particular de la superficie útil (*SUPERF*), el intervalo es muy amplio en los valores más bajos, lo que indica que este factor es poco representativo en los inmuebles más pequeños.

Figura 3.16. Relación de valor de la función de suavizado (s) con las variables precio de oferta y superficie útil, para el modelo de correspondencia en vivienda unifamiliar

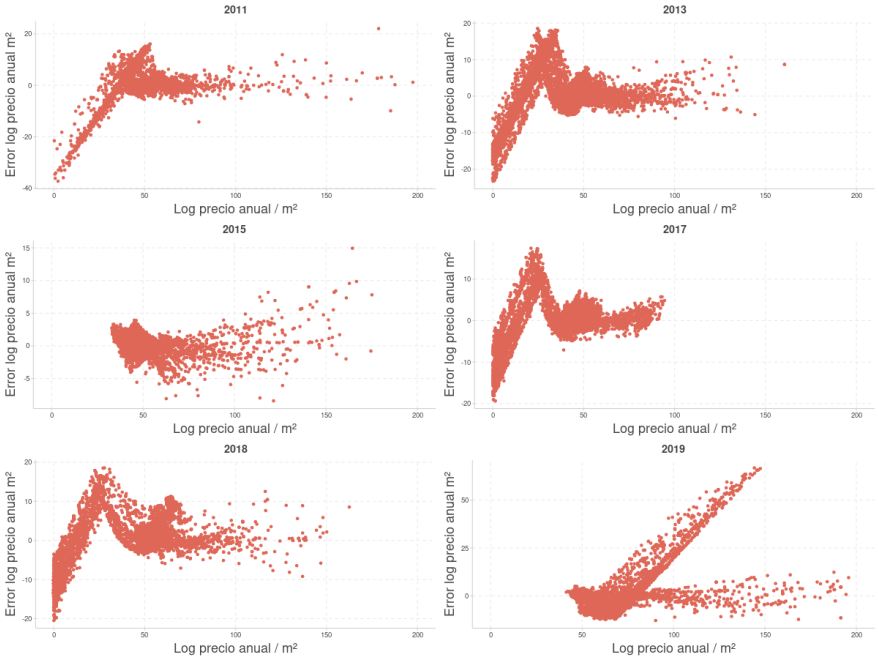


Fuente: elaboración propia.

Por otra parte, el precio de oferta (*preciom2_anualpred*) muestra una amplia variabilidad en todo el rango de valores, a excepción de los valores cercanos a 100 €/m²/año. En los valores superiores, el motivo podría ser la existencia de variables omitidas importantes como son el tamaño de la parcela, el estado de conservación de la vivienda, o la limitación de superficie útil a un máximo de 300 m². La incertidumbre en los valores bajos (menos de 100 m²) es menos importante porque para este tipo de propiedad son áreas infrecuentes.

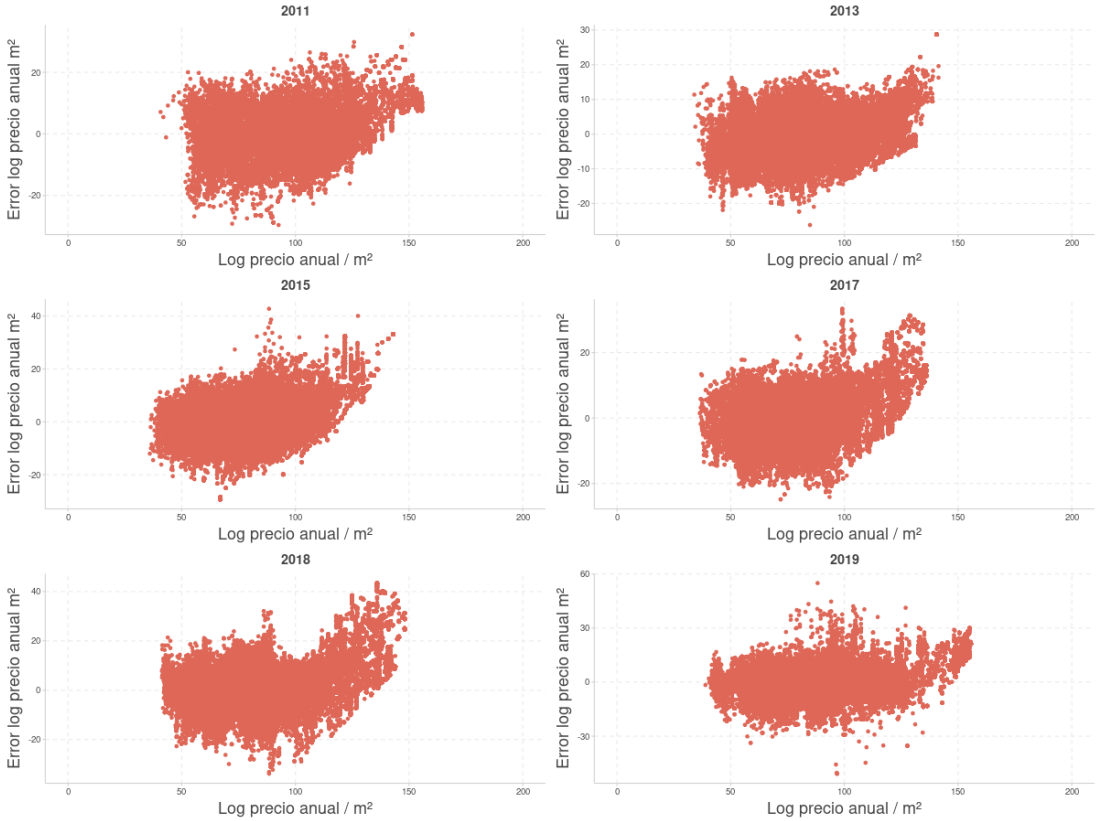
Para confirmar la hipótesis anterior, se representan los errores del modelo de viviendas unifamiliares en la Figura 3.17. Se aprecia como los errores muestran un patrón estable en el tiempo, excepto para los casos de 2015 y 2019 cuando ofrecen un patrón menos definido. En los primeros casos, los errores son mínimos para los precios más bajos, y que ascienden hasta un punto donde vuelven a descender y estabilizarse. Este comportamiento inestable de los residuos es habitual en muestras pequeñas y muy heterogéneas (Goh *et al.*, 2012), como la del segmento de estudio.

Figura 3.17. Residuos modelo de correspondencia en escala logarítmica, vivienda unifamiliar



Fuente: elaboración propia.

Figura 3.18. Residuos modelo de correspondencia en escala logarítmica, vivienda plurifamiliar

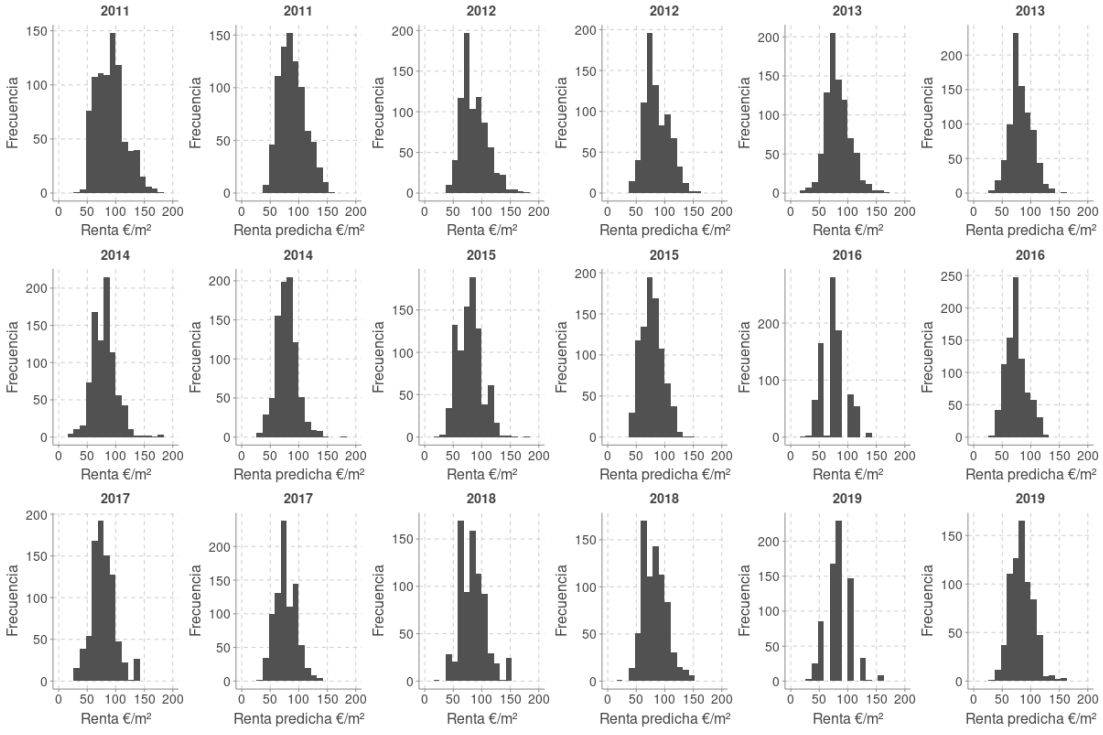


Fuente: elaboración propia.

Esto contrasta con los errores en el modelo para viviendas plurifamiliares de la Figura 3.18, donde los residuos, en escala logarítmica, están más cercanos a un patrón aleatorio. Además, la distribución de los valores son consistentes a lo largo del tiempo, y se aprecia que los errores son mayores en los precios más altos. Esto se debe a que los segmentos más caros contienen una mayor proporción de inmuebles singulares, cuyos precios no siguen el patrón general, y por tanto, el grado de imprecisión de los modelos en este tramo de precios es mayor.

Por último, debe comprobarse que la distribución de la variable de interés se mantiene en los ficheros donantes y receptores (Rässler, 2012). Para ello, primeramente, se evalúa si el modelo de imputación la EPF lo hace. En la Figura 3.19 se muestran las distribuciones del precio del alquiler original y estimada. Se observa que el modelo mantiene la forma de las distribuciones originales, con dos excepciones: en 2017 la predicción tiende a concentrar los valores en torno a la mediana; y en los años 2016 y 2019 los ficheros originales no ofrecen una forma continua cuando el modelo si lo hace.

Figura 3.19. Distribución original y predicciones para el modelo de imputación de valores de la EPF



Fuente: elaboración propia.

A tenor de lo anterior, la condición de preservación de las distribuciones marginales se considera válida al existir convergencia en los procesos de calibración.

3.4.3 Distribución conjunta

Para comprobar que distribución conjunta del proceso de correspondencia se mantiene, se debe asegurar que la distribución del modelo mantiene las propiedades del fichero de la EPF. En lo que se refiere a la distribución de frecuencias, existen diversas formas de medir el nivel de divergencia entre poblaciones, una de las más utilizadas (Leucescu y Agafitei, 2013) es la distancia Hellinger $H_d(P,Q)$, que puede aplicarse tanto para poblaciones continuas como discretas. Existen otras alternativas como las pruebas Chi-cuadrado, Kolmogorov Smirnov, Rao-Scott, Wald-Wolfowitz, que se estudian en detalle en Corder y Foreman (2014).

En este caso, la información de la EPF se conoce de forma discreta, agrupada por estratos (*YEAR*, *TAMAMU, TIPOEDIF*, *TIPOCASA*, *ZONARES*, *DENSI*, *ANNOCON*, *INTERINPSP*, *factorGASTOT6*, *NHABIT*, *CAPROV*), por tanto, se utiliza una variante de la distancia Hellinger aplicada a poblaciones discretas, calculada según la siguiente expresión analítica:

$$H_d(P,Q) = \frac{1}{n_d} \cdot \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2} \qquad [3.12]$$

donde n_d es el número de dimensiones, p_i es la probabilidad de inclusión del estrato i para la tabla de contingencia de los pesos del modelo, que en este caso serán los pesos calculados por la calibración, y q_i es la correspondiente probabilidad para el mismo estrato para la EPF. Los estratos i se engloban en un conjunto K total de estratos, cuyas probabilidades de inclusión p_i y q_i se definen como:

$$p_i = \frac{n_i}{N}, q_i = \frac{n'_i}{N'} \qquad [3.13]$$

donde N es el total de la tabla de contingencia, n_i la frecuencia tiene el estrato i, es decir el peso poblacional de este estrato en el conjunto P. Los n'_i y N' serían los respectivos valores para la EPF.

En este caso, se obtiene una distancia de Hellinger de 0,042 sobre un total de 2.723 estratos y 11 dimensiones, con una media de 248 estratos por dimensión. Según Leucescu y Agafitei (2013) se considera que esta distancia representa que dos distribuciones similares cuando su valor es menor de 0,05[37]. Se puede concluir, por tanto, que se mantiene la distribución conjunta de pesos poblacionales, al ser la distancia del mismo orden de magnitud que las obtenidas por los mismos

[37]Es cierto que esta medida debe usarse con precaución porque no tiene en consideración la variabilidad debida al diseño del muestro o cuando existen un gran número de categorías.

autores (Leucescu y Agafitei, 2013) en las armonizaciones de las encuestas de condiciones de vida (EU-SILC) y de población activa europea (EU-LFS). En cuyos casos obtuvieron un valor cercano al 0,04.

Para la distribución conjunta de los precios, se ha comprobado que los parámetros principales del fichero de la EPF se mantienen en el fichero definitivo, tanto en órdenes de magnitud como en tenencia. La Tabla 3.12 muestra estos parámetros en ambos ficheros (media, cuantiles y desviación estándar ponderados[38]). Se observa que la desviación típica en el fichero definitivo es ligeramente menor, a excepción del año 2014 cuando se produce un fuerte aumento de la desviación. Los valores absolutos de la media y cortes de cuantiles son superiores debido al efecto del suavizado exponencial aplicado a la EPF.

Tabla 3.12. Parámetros sobre precio €/m²/año población original y final

	Fichero EPF					Fichero final				
Año	**Media**	**Dev**	**Q1**	**Q2**	**Q3**	**Media**	**Dev**	**Q1**	**Q2**	**Q3**
2011	92	657	71	90	106	100	599	81	101	119
2012	88	569	72	83	103	96	609	76	99	115
2013	84	441	70	80	97	91	474	74	93	109
2014	81	507	68	80	91	90	803	71	87	106
2015	79	435	63	80	90	86	398	72	87	101
2016	76	437	58	75	88	83	375	71	82	98
2017	79	442	58	75	95	86	417	74	86	100
2018	83	559	62	84	95	90	562	72	89	105
2019	85	512	70	87	103	95	549	76	96	110

Fuente: elaboración propia

Aún cuando los el nivel de ajuste es muy alto, se puede comprobar gráficamente que las medidas promedio de la variable de interés varían sensiblemente en función de los elevadores muestrales utilizados, a pesar de referirse al mismo colectivo. Lo cual puede comprobarse en la Figura 3.20, dónde se representa la suma de los precios ponderados de los estratos individuales[39] usando dos conjuntos de pesos: los originales de la EPF y los estimados por el proceso de calibración ($g \cdot w$).

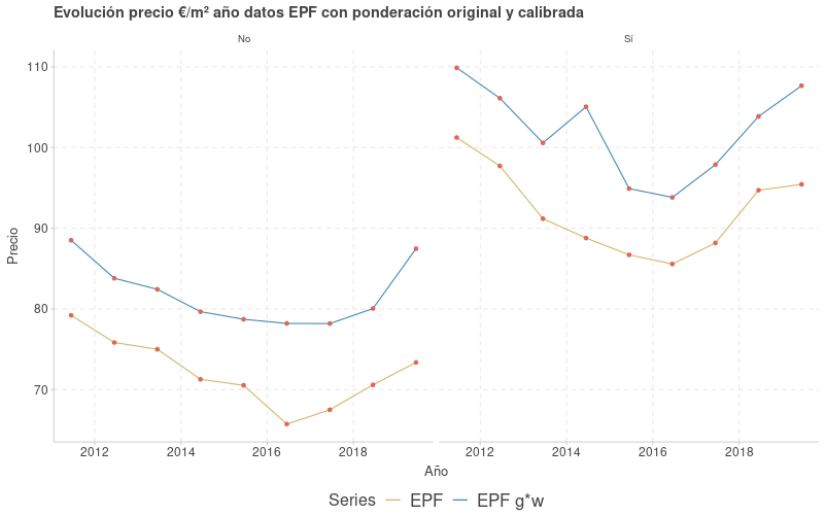
La Figura muestra que la serie de precios basada en la calibración tiene comportamientos anómalo, desde un punto de vista de lógica de mercado. Por

[38] En cada fichero se usan pesos diferentes, para la EPF los pesos del fichero original y en el fichero de oferta se usan los factores de elevación procedentes de la calibración.

[39] Se parte de las celdas de menor tamaño siguiendo la estratificación de la población mediante las covariables del modelo de correspondencia.

ejemplo, se observa un incremento importante de precios entre 2018 y 2019 para el resto de Comunidad de Madrid, y una subida puntual del precio en 2014 paras la ciudad de Madrid, ambos valores no se corresponden a ninguna causa de mercado justificable.

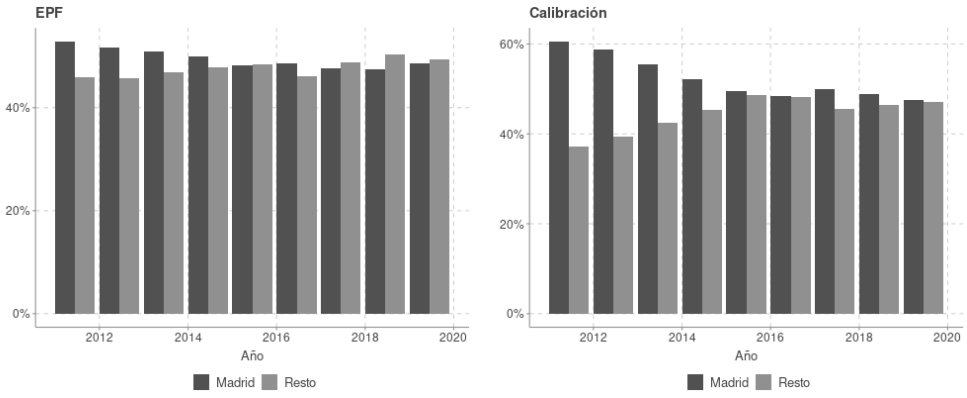
Figura 3.20. Series de precio promedio: pesos EPF y pesos calibrados



Fuente: elaboración propia.

Las diferencias en las tendencias de la figura anterior se deben a un efecto de composición con los nuevos pesos, que lógicamente difieren de los originales. La Figura 3.21[40] muestra las diferencia entre los pesos originales y los calibrados para los estratos definidos.

Figura 3.21. Pesos poblacionales EPF y calibración por Madrid o resto de zonas



Fuente: elaboración propia.

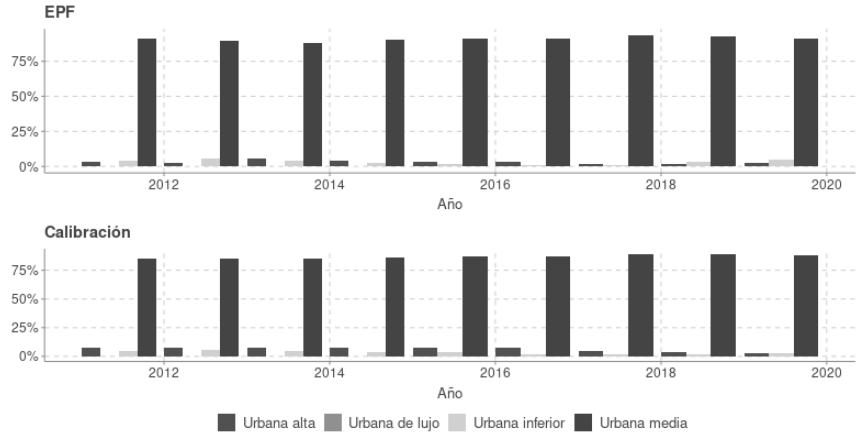
Los pesos de la EPF son relativamente estables a lo largo del tiempo. En cambio, los pesos calibrados parten de una situación mucho más desequilibrada en 2011

[40] La variable *CAPROV* indica si la observación se encuentra en la capital de provincia.

que en el 2019.

Para otros criterios de estratificación, como el tipo de zonas residencial, no se aprecian unas diferencias tan acusadas. La Figura 3.22 que muestra ambas distribuciones de pesos, indica que existe una mayor representación de las zonas minoritarias en la calibración, pero la desigualdad es mínima.

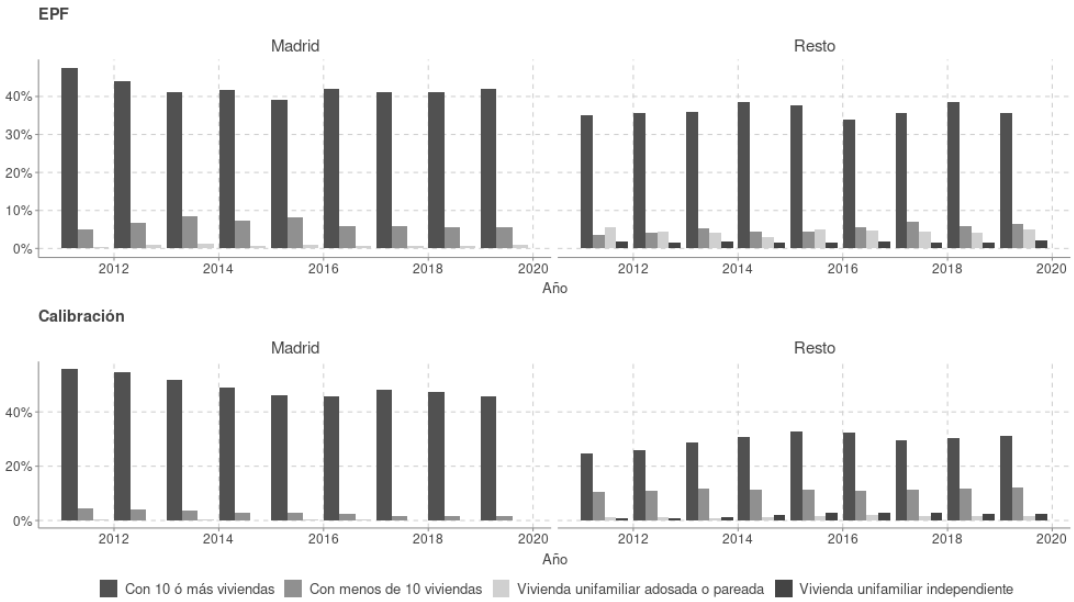
Figura 3.22. Pesos poblacionales EPF y calibración por tipo de zona residencial



Fuente: elaboración propia.

En el caso del desglose por edificio, hay una mayor distribución por tipo en la ciudad de Madrid, y una mayor diversidad de tipos en el resto de provincia en la población calibrada por el censo, como se observa en la Figura 3.23.

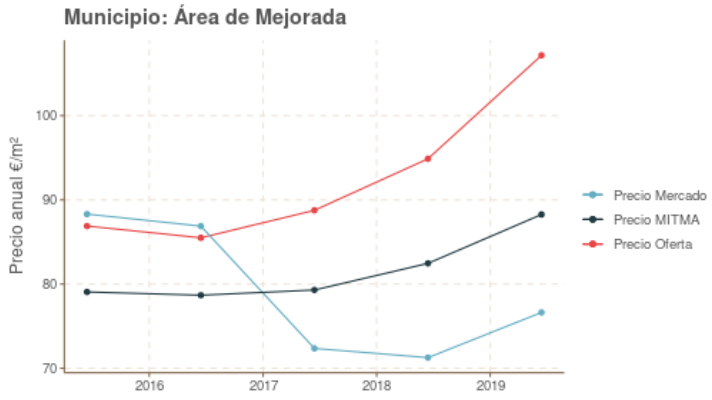
Figura 3.23. Pesos poblacionales EPF y calibración por tipo de edificio



Fuente: elaboración propia.

Las medidas anteriores no tienen en cuenta el desglose zonal, ya que el dato a este nivel no está disponible en el fichero “donante” de la EPF y no puede utilizar como variable común para hacer el enlace. Se puede anticipar que la falta de referencias geográficas genera sesgos en las estimaciones de precios de zona en el modelo, a la vista de los valores del ejemplo de la Figura 3.24, para las series de precios en el área de Mejorada del Campo (zona en la que se aprecia de forma muy acusada este efecto). En la cual se observa una caída del precio de mercado entre 2015 y 2018, que no guarda coherencia con el precio registrado oficialmente por MITMA en ese municipio.

Figura 3.24. Precios para vivienda plurifamiliar, área de Mejorada del Campo



Fuente: elaboración propia.

Se puede apreciar también que la misma inconsistencia con MITMA se produce con los precios de oferta. El motivo de la misma podría deberse a que el modelo de mercado no ha tenido en cuenta la zona geográfica específica, al contrario que en oferta o MITMA. Por tanto se puede asumir que se introduce un sesgo por una insuficiente especificación zonal de los modelos hedónicos de mercado (al no disponerse de esta información en la EPF).

Para comprobar lo anterior, la Tabla 3.13 muestra dos métricas de discrepancia entre las series de oferta y mercado con respecto a las de MITMA en todas las zonas, entre 2015 y 2019. La primera medida (divergencia) compara el signo de las variaciones de cada una de las series, y expresa el número de veces en los que el signo de la variación difiere (normalizado por el número de observaciones). La segunda métrica (diferencia) representa la desviación media en porcentaje, entre las variaciones de la serie y la de MITMA.

Tabla 3.13. Divergencia variación anual de precios respecto a MITMA

		% divergencia		% diferencia	
Tipo	**N**	**Oferta**	**Mercado**	**Oferta**	**Mercado**
Plurifamiliar	668	7,2%	32,0%	5,1%	27,9%
Unifamiliar	236	40,3%	45,3%	12,2%	41,6%

Fuente: elaboración propia

Los resultados muestran que se puede generalizar la anomalía observada en Mejorada del Campo para el resto de las zonas, donde existe una divergencia alta en las series de mercado generadas por el modelo (superiores al 30%). En cambio, la coincidencia para la oferta es muy alta en términos de signo, especialmente para las viviendas plurifamiliares, con una divergencia del 7,1% y una desviación en términos absolutos de un 5,1%, lo que hace suponer que estas series están altamente correlacionadas.

Las viviendas unifamiliares muestran un comportamiento más irregular, siendo las tasas de coincidencia en signo altas en todos los casos.

La causa principal del comportamiento anterior procede de la heterogeneidad espacial en las distribuciones de precios, que por otra parte, es un fenómeno conocido y ampliamente documentado. Entre otros autores, la cuestión ha sido analizada en profundidad por Hu (2022), Wu (2020), Helbich (2014), Páez (2008) y Kestens (2006). Para nuestro caso, se pueden identificar dos orígenes de la heterogeneidad:

- La calibración final con la EPF no tiene información zonal.
- Ninguno de los modelos de mercado se refieren a zonas concretas, sino que son estratos de tipo funcional o de características[41].

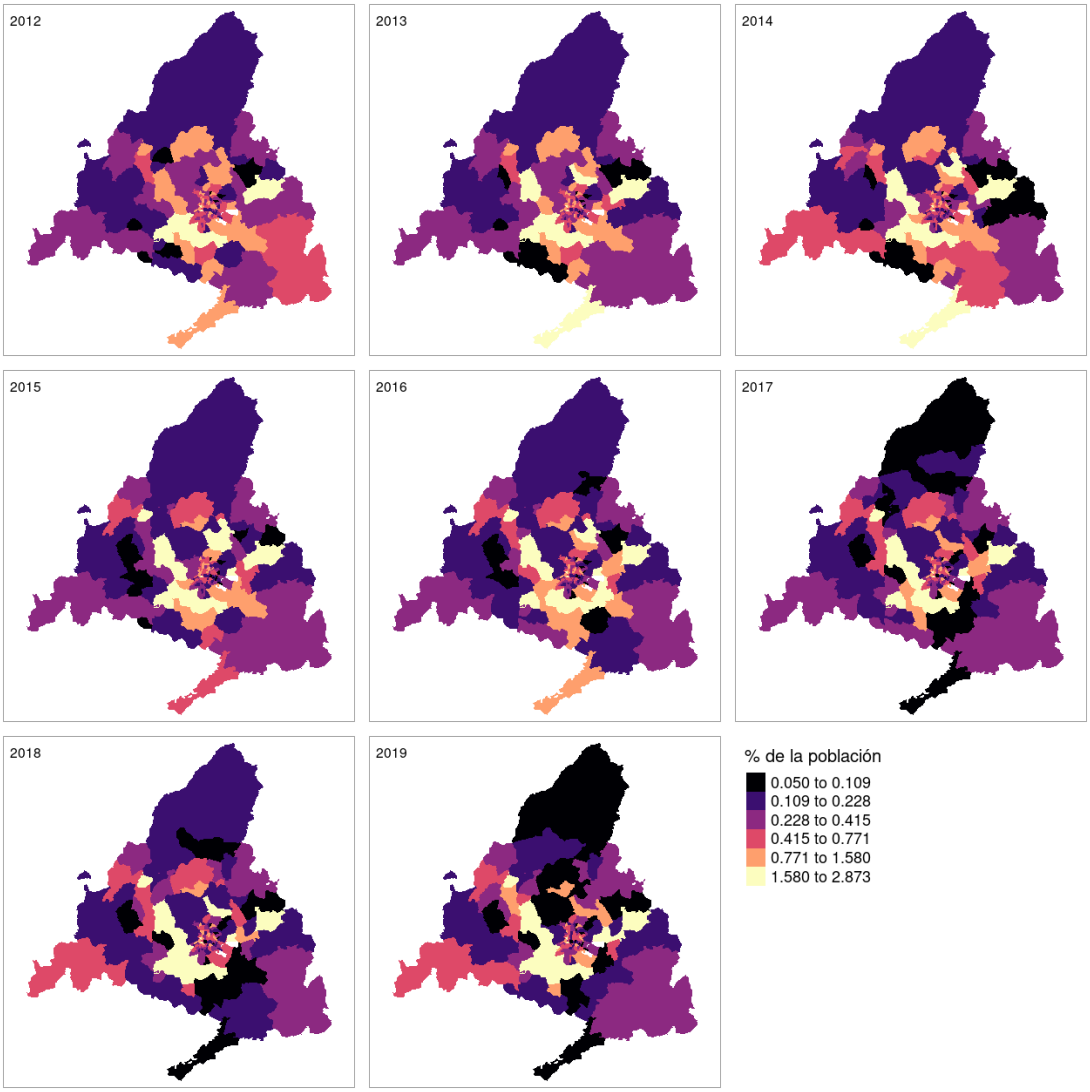
[41] Incluso las características de clasificación de zona: rural, urbana, densamente poblada o según ingresos no recogen las diferencias entre zonas

3.4.4 Distribución espacial de la población

Es requisito indispensable que la distribución de los pesos muestre estabilidad temporal, y se corresponda la de la poblacional real. De la misma manera, y aunque se parta de la limitación del desconocimiento de las distribuciones exactas del alquiler por zona geográfica, el proceso de calibración ideal debería ser capaz de replicar la distribución zonal de oferta en la medida de lo posible[42].

En el epígrafe anterior, los niveles de ajuste del modelo confirman que el modelo replica el desglose funcional de la muestra. Por tanto, si es necesario un ajuste, será para garantizar que se mantiene la coherencia zonal de la información, manteniendo el comportamiento funcional actual.

Figura 3.25. Distribución espacio-temporal de los pesos poblacionales en vivienda plurifamiliar, toda la Comunidad de Madrid



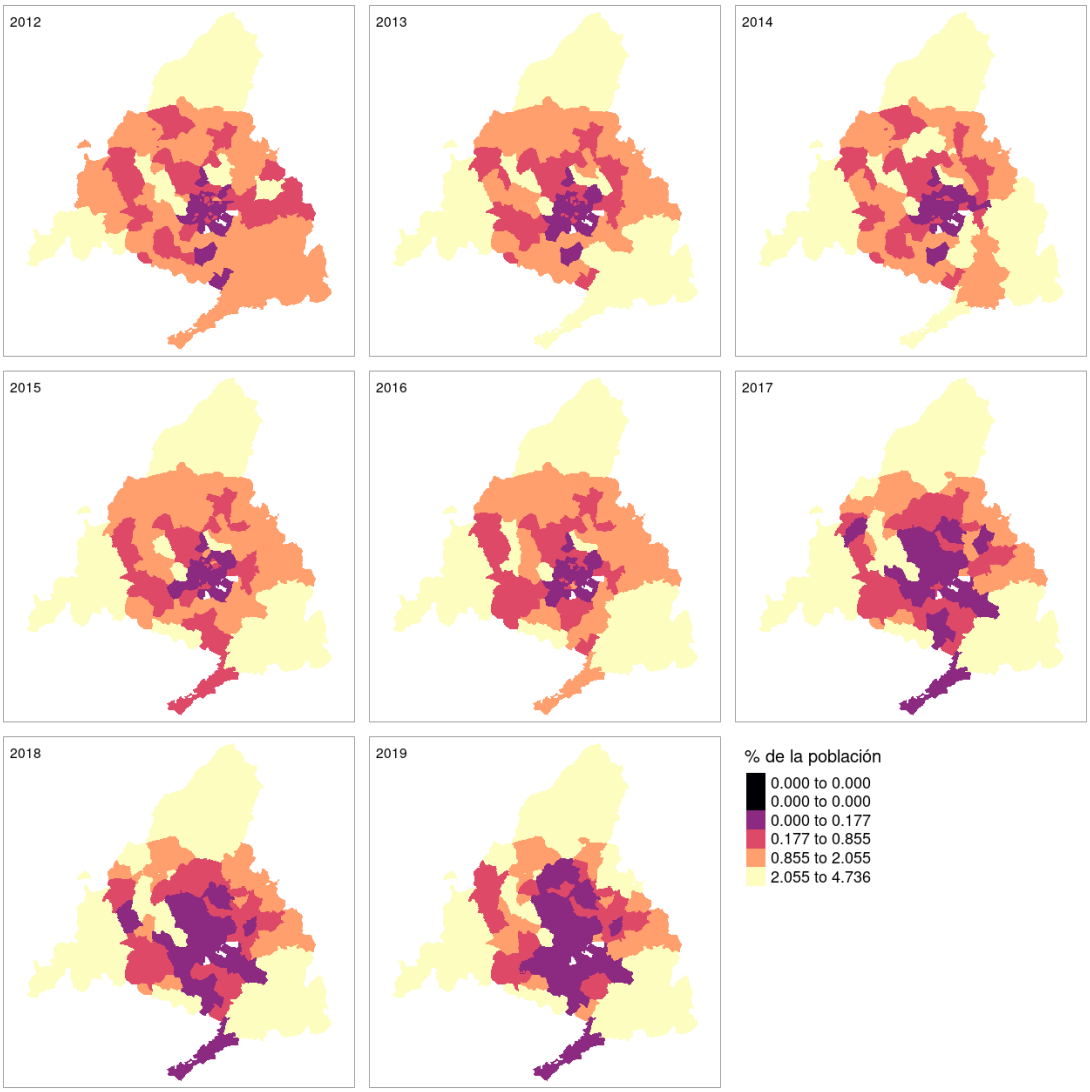
Fuente: elaboración propia.

[42] Entendiendo que en ausencia de información de las rentas reales, el dato de oferta ofrece una medida aproximada, aunque sea en términos de orden de magnitud y evolución temporal.

Para las viviendas plurifamiliares, se observa en la Figura 3.25, que la representación de las zonas rurales decrece en el tiempo, mientras que, la zonas metropolitanas sur y oeste ganan progresivamente importancia en la muestra. Existe una mayor variabilidad en el tiempo en las zonas centrales de la Comunidad, con respecto a las zonas exteriores.

Las viviendas unifamiliares, en cambio, se concentran en la corona justamente exterior al centro, ampliándose el radio interior de forma progresiva, como vemos en la Figura 3.26. Esto indica que la población de alquiler de las zonas metropolitanas exteriores ganan progresivamente más importancia.

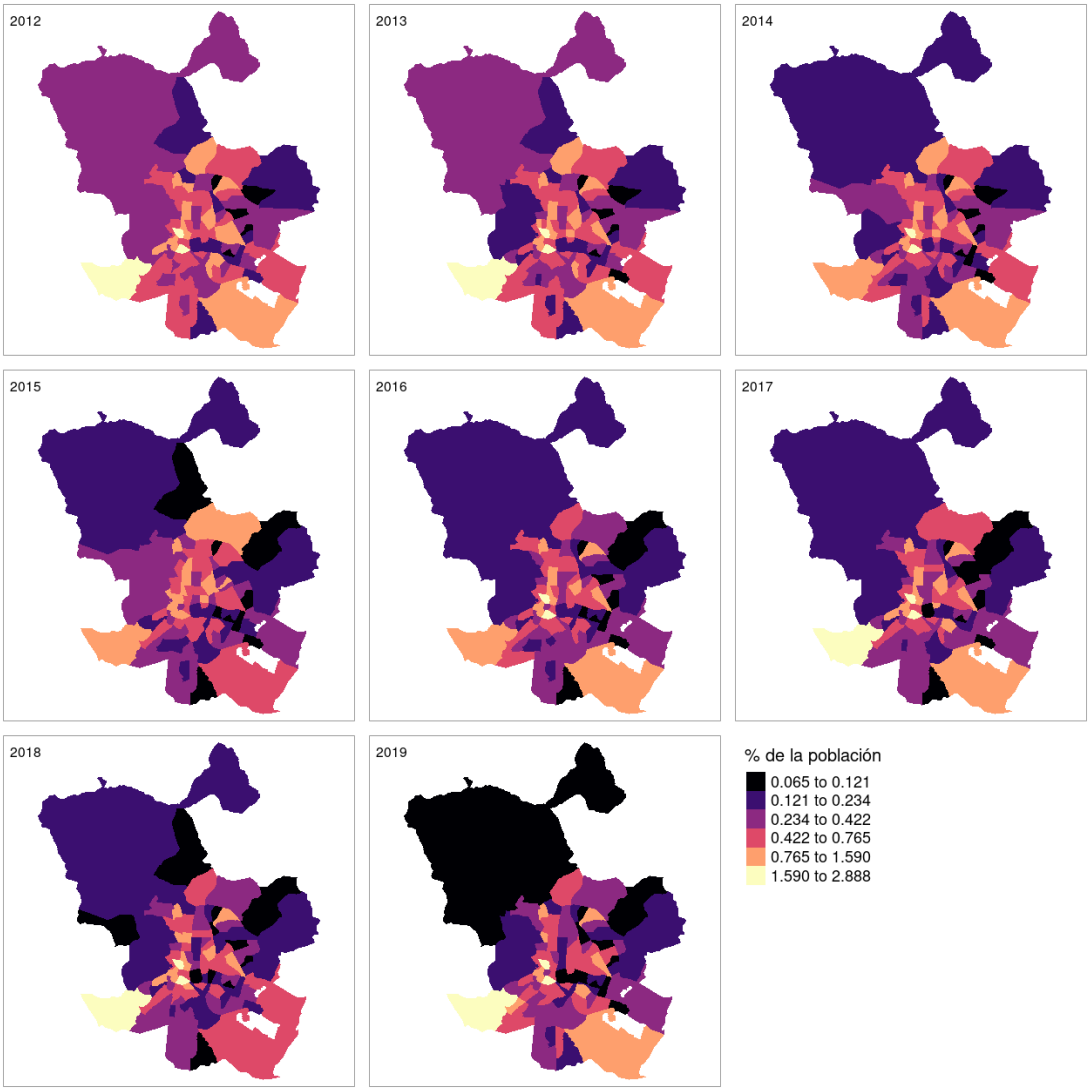
Figura 3.26. Distribución espacio-temporal de los pesos poblacionales en vivienda unifamiliar, toda la Comunidad de Madrid



Fuente: elaboración propia.

Para el caso de la ciudad de Madrid, se observa que el peso de las distintas zonas varía en el tiempo, como se aprecia en la Figura 3.27, aunque son la zona central y la sur las que mantienen más registros en términos relativos.

Figura 3.27. Distribución espacio-temporal de los pesos poblacionales para todos los tipos, ciudad de Madrid



Fuente: elaboración propia.

La Tabla 3.14 muestra las zonas que más han variado en términos relativos. Destaca la zona metropolitana del noroeste como aquella con variaciones más altas a partir del año 2017. Por ejemplo, Majadahonda pasa de un 1,1% del total de la población al 4,8%. En general, estos municipios del noroeste (Majadahonda, Las Rozas, Collado Villalba y Pozuelo de Alarcón) representaban un 4,2% en 2011, mientras que en 2019 cuentan con un 14,4% de las viviendas en alquiler. Estas áreas conforman el eje de mayores ingresos de la Comunidad, y también, cuentan con un nivel de precios por superficie mayor a la media.

Tabla 3.14. Zonas con mayor variación en porcentaje

Zona	2011	2012	2013	2014	2015	2016	2017	2018	2019
Majadahonda	1,1%	1,4%	1,6%	2,0%	2,6%	2,7%	4,3%	4,7%	4,8%
Rozas de Madrid, Las	1,1%	1,4%	1,7%	2,1%	2,7%	2,7%	4,2%	4,5%	4,5%
Coslada	0,9%	1,1%	1,3%	1,5%	2,3%	2,4%	3,8%	3,9%	4,1%
Pozuelo de Alarcón	1,0%	1,3%	1,4%	1,8%	2,5%	2,3%	3,5%	3,7%	3,7%
Collado Villalba	1,0%	1,2%	1,3%	1,5%	2,3%	2,3%	3,6%	3,6%	3,4%
Alcalá de Henares	2,9%	3,1%	3,1%	3,3%	2,8%	2,6%	2,2%	2,2%	2,2%
Móstoles	2,0%	1,9%	2,0%	1,9%	1,9%	1,9%	1,9%	1,9%	2,4%
Getafe	2,0%	2,0%	1,9%	1,7%	1,8%	1,8%	1,5%	1,8%	1,6%
Embajadores	2,1%	1,9%	2,0%	1,8%	1,6%	2,2%	2,1%	2,4%	2,4%
Universidad	1,7%	1,8%	1,7%	1,8%	1,3%	1,7%	1,7%	2,0%	1,8%

Fuente: elaboración propia

Las diferencias por tipología y capital de provincia (Tabla 3.15) muestran mayor estabilidad en Madrid (mediana del 0,11% contra un 0,19%). También existen menores diferencias en la vivienda plurifamiliar que en la unifamiliar (del 0,82% y un 0,15% respectivamente, para el resto de la CAM). Más concretamente, los cambios más importantes y fuente de variabilidad en los precios, se producen en viviendas unifamiliares fuera de la capital.

Tabla 3.15. Zonas con mayor variación anual media (periodo 2011 a 2019)

Capital	Tipo	Mín.	p05	p25	p50	p75	p95	Máx.
No	Todos	0,00%	0,02%	0,08%	0,19%	0,60%	2,61%	3,63%
	Plurifamiliar	0,01%	0,01%	0,06%	0,15%	0,57%	2,59%	3,85%
	Unifamiliar	0,01%	0,06%	0,32%	0,82%	1,34%	4,05%	10,74%
Sí	Todos	0,00%	0,01%	0,06%	0,11%	0,20%	0,31%	0,58%
	Plurifamiliar	0,00%	0,01%	0,06%	0,11%	0,20%	0,31%	0,56%
	Unifamiliar	0,00%	0,00%	0,00%	0,02%	0,13%	0,66%	1,31%

Fuente: elaboración propia

A lo largo de este capítulo se ha desarrollado un modelo de mercado que relaciona los precios de oferta con los de mercado, que sin embargo cuenta con limitaciones por la débil especificación zonal de la EPF. Esta cuestión se solventará con el modelo hedónico del capítulo 6, incorporando información zonal a través de medidas de las accesibilidad presentadas en el próximo capítulo.

Anexo 3a. Métodos de calibración

Aunque originalmente la calibración se orientó a la reducción de la varianza, se utiliza a menudo para corregir los sesgos muestrales. En contraposición con el enfoque clásico, en el que se construyen grupos de respuesta homogéneos[43], en la calibración, las variables de respuesta deben conocerse únicamente para las unidades que responden, junto con los totales poblacionales, lo que resulta bastante restrictivo. Para resolverlo, Deville desarrolló la teoría de calibración "supergeneralizada" (Deville, 2000), donde no es necesario conocer los totales poblacionales de las variables en las subpoblaciones sin respuesta, y se trabaja solo con las unidades cuya respuesta es conocida.

El método supone utilizar información auxiliar relacionada con la variable de estudio para ajustar los pesos de la muestra. Su base consiste en que, dada la característica Y de estudio, existe una serie de variables auxiliares X fuertemente relacionadas con Y, cuyos datos son conocidos o son fácilmente accesibles.

La calibración permite mejorar los estimadores de los parámetros poblacionales basados en técnicas tradicionales calculadas como métodos lineales, cuadráticos o por diseños muestrales complejos. Todo ello redunda en métodos con menor error, y permite trabajar con poblaciones con un menor tamaño. Deville y Särdnal (1992) demuestran, además, que cuanto más fuerte es la relación de las variables auxiliares con las variables de estudio, más precisa es la calibración.

La idea de la calibración parte de que dada una función de distancia d, que se establece entre los pesos iniciales y finales, se puede encontrar una razón a aplicar a los pesos originales π_k, denominada g_k, que cumple que la suma de sus distancias es mínima. La condición a minimizar por tanto sería:

$$\sum_{R_k=1}^{N} d(g_k, 1) \qquad [3.14]$$

donde las variables y_i son las variables de interés conocidas, tanto de forma agregada como a nivel individual. La condición anterior está sujeta a las restricciones de calibración:

$$\sum_{k=1}^{N} y_i = \sum_{R_k=1}^{N} \frac{g_i}{\pi_k} \cdot y_i \qquad [3.15]$$

De una forma más formal, dada una población U compuesta de N elementos

[43] La muestra se divide en celdas donde se supone que la distribución de la variable de la respuesta es uniforme. Por tanto, la tasa de respuesta es una estimación máxima en verosimilitud para esa distribución.

(k) distintos, identificados a través de sus etiquetas $i = 1, ..., N$, se parte de una serie de características de interés y_i asociadas con el elemento i que se conoce exactamente, y sin error, observando el elemento i.

Una muestra s es un subconjunto de U cuyos valores asociados de Y, identificados como $\{(k, y_k)\}$, que se seleccionan según con un diseño muestral específico que asigna una probabilidad conocida $p(s) > 0$ para todo $s \in S$. Siendo S el conjunto de las todas las posibles muestras s, y que cumple ques $\sum_{S \in S} p(s) = 1$. El total poblacional T de la variable Y se calcularía como:

$$T_Y = \sum_{k \in U} y_k \qquad [3.16]$$

La muestra cuenta con un vector de variables auxiliares $X = (X_1, ...X_J)$ que es perfectamente conocido, para todos los elementos de la población U. De tal forma que consideramos el estimador de Horvitz-Thompson[44] $\hat{T}_Y$ asociado como:

$$\hat{T}_{Y_\pi} = \sum_{k \in S} d_k \cdot y_k \qquad [3.17]$$

que pretende es modificar los pesos originales d_k, calculados como:

$$d_k = \frac{1}{\pi_k} \qquad [3.18]$$

por otros pesos ω_k, de forma que, el estimador basado en dichos pesos proporcione estimaciones perfectas para X, es decir, que cumpla:

$$\sum_s \omega_k X_k = T_X = (T_{X_1}, .., T_{X_J}) \qquad [3.19]$$

y estén tan próximos como sea posible, según una medida de distancia dada, a los pesos originales d_k. El método más común para definir esta distancia es la suma ponderada de los cuadrados de las distancias:

$$\sum_{k \in S} \frac{(\omega_k - d_k)^2}{q_k \cdot d_k} \qquad [3.20]$$

donde q_k son constantes positivas, que se resuelven como un problema de minimización de la expresión anterior, con la siguiente restricción:

[44] Un estimador de Horvitz-Thompson es un método para estimar el total y la media de una pseudopoblación en una muestra estratificada (Fuller, 2011; Särndal *et al.*, 2003).

$$\sum_{s} \omega_k \cdot X_k = T_X \qquad [3.21]$$

usando el método de los multiplicadores de Lagrange, se obtienen los siguientes pesos, ya calibrados como:

$$\omega_k = d_k + d_k \cdot q_k \cdot \lambda X'_k \qquad [3.22]$$

suponiendo que la inversa de $T_S = \sum_{k \in S} d_k \cdot q_k \cdot X_k \cdot X'_k$ existe, el estimador calibrado vendría dado como el estimador general de regresión, véase (Cassel *et al.*, 1976), definido según la expresión analítica:

$$\hat{T}_{Y_{reg}} = \sum_{k \in S} \omega_k y_k = \hat{T}_{Y_\pi} + (\hat{T}_X - \hat{T}_{X_\pi}) \cdot \hat{B}_S \qquad [3.23]$$

La forma que el total estimado, $\hat{T}_{Y_{reg}}$, dependerá del diseño muestral y de las constantes q_k elegidas. Si se trabaja con una única variable auxiliar $X = X_1$, y $q_k = \frac{1}{x_k}$ y se realiza un muestreo aleatorio simple, entonces el total se calcula como un estimador de razón:

$$\hat{T}_{Y_{reg}} = \frac{\overline{x}}{\overline{y}} \cdot X \qquad [3.24]$$

donde $\overline{y}$ y $\overline{x}$ son las correspondientes medias muestrales de la variable Y y la variable X. El estimador $\hat{T}_{Y_{reg}}$ no es generalmente insesgado pero los pesos ω_k serán muy próximos a d_k, por lo que es asintóticamente insesgado (Särndal, 2007).

Existen alternativas para la construcción de los estimadores de calibración que no modifican la medida de distancia, sino que cambian el proceso de construcción del estimador a través de la modificación de alguna de estas dos condiciones (y en muchos casos, apoyándose en modelos de superpoblación):

- Minimización de una distancia.
- Que los pesos equilibrados den estimaciones perfectas para las variables auxiliares.

Martínez (2002) menciona cuatro tipos de estimadores de calibración, entre los que el más flexibles es el asistido por modelos:

- Estimadores de calibración para una familia de distancias.
- Estimadores de calibración basados en una forma funcional.
- Estimadores de calibración cosméticos.
- Estimador de calibración asistido por modelos.

Wu (2001) clasifica los métodos asistidos por en tres tipos:

- Estimadores de regresión generalizada (GREG) (Cassel *et al.*, 1976) y (Särndal, 1980).
- Estimadores de calibración (Deville y Särndal, 1992).
- Empíricos de probabilidad (Chen y Qin, 1993; Chen y Sitter, 1999).

Los estimadores de calibración asistidos por modelos se construyen al sustituir en el proceso de calibración la restricción de la expresión [3.25] por otra más adecuada, ya que se asume implícitamente un modelo lineal de regresión, entre la variable de estudio Y y las variables del vector X, en la población de estudio:

$$\sum_{k \in S} \omega_k \cdot y_k = T_X \qquad [3.25]$$

La relación entre ambas variables no tiene porque acomodarse a una regresión lineal simple, de ahí la necesidad de plantear una mecanismo más flexible que tome una función de relación que se adapte a cada situación. Para ello se puede utilizar un modelo sobre un estimador de regresión generalizada, denominada GREG (Deville y Särndal, 1992; Särndal *et al.*, 2003). Este tipo de calibración, asume que la relación entre las variables Y y el vector X se describe por un modelo de superpoblación, especificado como:

$$\begin{aligned} E_\xi(y_k|X_k) &= \mu(X_k, \theta) \\ V_\xi(y_k|X_k) &= v_k^2 \cdot \sigma^2 \end{aligned} \qquad [3.26]$$

Con $k = 1, 2, ..., N$, donde $\theta = (\theta_0, \theta_1, ...\theta_J)'$ y σ^2 son parámetros poblacionales desconocidos, $\mu(X, \theta)$ es una función conocida de X y θ, v_k es una función conocida de X_k o bien de $\mu_k = \mu(X_k, \theta)$ y $E\xi$ y $E\xi$ son, respectivamente, la esperanza y la varianza con respecto al modelo de superpoblación.

Esta especificación general incluye tres de los casos más comunes de calibración con modelos que son: los modelos de regresión lineales, los no lineales y los generalizados.

Los enfoques anteriores se basan en un contexto de un modelo de regresión e incorporan esencialmente las variables auxiliares a través de sus medias poblacionales conocidas, incluso cuando se conocen las variables auxiliares para cada unidad en el población. Siendo $1/\pi_k$ los pesos ordinarios del muestreo para la observación k-ésima, dónde π_k es la probabilidad de inclusión de k:

$$V_\xi(y_k|X_k) = v(\mu_k)\ k = 1, 2, ...N \qquad [3.27]$$

donde $\mu_k = E_\xi(y_k\ X_k)$ es una función de enlace y V es la función varianza. El estimador de calibración asistido por un modelo T_Y se define como:

$$\hat{T}_Y = \sum_{k \in S} \omega_k\ y_k \qquad [3.28]$$

donde los pesos calibrados w_k son mínimos, según una medida de distancia con respecto a d_k, estando $\hat{T}_Y$ sujeto a la condición:

$$\sum_{k \in S} \omega_k \mu(X_k \hat{\theta}) = \sum_{k=1}^{N} \mu(X_k \hat{\theta}) \qquad [3.29]$$

una vez minimizada la medida de distancia, se obtiene el siguiente estimador:

$$\hat{T}_Y^* = \hat{T}_{Y_\pi} + \left(\sum_{k=1}^{N} \hat{\mu_k}\ -\ \sum_{k \in S} d_k \hat{\mu_k} \right) \ddot{B}_N^* \qquad [3.30]$$

donde:

$$\ddot{B}_N^* = \frac{\sum_{k \in S} d_k \cdot q_k \hat{\mu}_k \cdot y_k}{\sum_{k \in S} d_k \cdot q_k \hat{\mu}_k^2} \qquad [3.31]$$

Wu (2001) identifica las propiedades más importantes del estimador de calibración:

1. Bajo ciertas condiciones, tanto $\hat{T}_Y$ como $\hat{T}_Y^*$ son iguales a $\hat{T}_{Y_\pi} + O\left(n^{-1/2}\right)$ y son asintóticamente insesgados para T_Y, con respecto al diseño, sin tener en cuenta si el modelo es correcto o no.
2. Si $q_k = 1/v_k^2$, entonces $\hat{T}_Y$ y $\hat{T}_Y^*$ pueden ser estimadores de calibración basados modelos, tanto de tipo lineal como no lineal. Es decir, ambos son consistentes respecto al diseño, independientemente del tipo de modelo. En el caso de un modelo sin error, es decir, $y_k = \mu_k$, la condición puede expresarse como:

$$\hat{T}_Y = \hat{T}_Y^* = \hat{T}_Y \qquad [3.32]$$

Además, si se usa un modelo lineal ambos, $\hat{T}_Y$ y $\hat{T}_Y^*$, se reducen al estimador convencional de calibración (Deville y Särndal, 1992).

Si la encuesta parte de una muestra de tamaño n, donde w es el vector de pesos originales, de dimensión $n \times 1$, y ω' el vector homólogo de pesos transformados, cualquier procedimiento de reponderación que se aplique dará lugar a una relación funcional del tipo $\omega' = \omega'(\omega, X)$. De manera que los nuevos pesos van a ser función de los originales y de las variables auxiliares elegidas.

La información auxiliar proporcionada por la encuesta va a estar contenida en una matriz $X_{n \cdot p}$ donde en cada fila aparecen los valores de las variables auxiliares para cada individuo de la muestra. Los nuevos pesos han de cumplir la condición de equilibrado de la muestra, es decir, $X'\omega' = x$, siendo x el vector de efectivos poblacionales proporcionados por las fuentes externas utilizadas. Con los pesos ω', se calcularían las nuevas estimaciones para cualquier variable Y de interés.

Medidas de distancia en la calibración

La calibración parte de la definición previa de una función de distancia $G(\omega, \omega')$ entre los pesos originales ω y los calibrados ω', para la que se exige que:

$$\sum_{k=1}^{n} \omega_k \cdot G(\omega_k \cdot \omega'_k) = N \qquad [3.33]$$

sea mínimo para el conjunto de la muestra, con la restricción:

$$\sum_{k=1}^{n} \omega'_k = N \qquad [3.34]$$

Es decir, que la suma de pesos transformados debe recuperar un determinado total de población. Siendo h el cociente entre ponderaciones w'_k/w_k, se definen las dos familias de distancias más usualmente utilizadas:

- Cuadrática: $G(h) = \left(\frac{h-1}{2}\right)^2$.
- Logarítmica $G(h) = h \;\; log(h) - h + 1, h > 0$.

Tabla 3.16. Medidas de distancia básicas para calibración

Medida de distancia	Especificación
Chi cuadrado	$(\omega - d)^2/2qd$
Chi cuadrado modificado	$(\omega - d)^2/2q\omega$
Mínima entropía	$q^{-1}(-d\log(\omega/d) + \omega - d)$
Mínima entropía modificada	$q^{-1}(w\log(\omega/d) - \omega - d)$
Hellinger	$2(\sqrt{\omega} - \sqrt{d})^2/q$

Existen múltiples forma de especificar la medida de distancia $D(\omega, d)$. Deville y Särndal (1992) recogen 5 medidas de distancia, mostradas en la Tabla 3.16.

Es importante señalar que, dependiendo de la función de distancia elegida $D(\omega, d)$, puede que no exista una solución analítica para la condición:

$$\frac{\partial Q}{\partial \omega_i} = \frac{(\omega_i - d_i)}{q_i d_i} - \lambda x_i \qquad [3.35]$$

y es posible que requiera una aproximación numérica de w_i usando Newton-Raphson o un método similar. Además, la solución a la ecuación [3.35] puede producir ponderaciones positivas, negativas o extremadamente grandes que pueden ser no ser deseables en un contexto de muestreo.

En términos de eficiencia, Deville y Särndal (1992) demostraron que para muestras de tamaño medio y grande, la elección de la función $D(\omega, d)$ no tiene un gran impacto en la varianza del estimador elegido. Deville y Särndal también demostraron que bajo ciertas condiciones, el estimador es asintóticamente equivalente a GREG para cualquier función de distancia $D(w, d)$. Por lo tanto, la elección de la función de distancia no es importante para muestras grandes, sino que depende del esfuerzo de proceso de resolver la condición de la expresión [3.35].

Métodos de calibración

Asociados a estas funciones de distancia, existen diversos métodos de calibración, que proponen funciones de transformación de los pesos nuevos respecto a los originales. Entre los más comunes, se encuentran los métodos lineal, exponencial y lineal *logit* truncado.

La transformación lineal se basa en la medida de distancia Chi-cuadrado, y es la más comúnmente aplicada por su sencillez y porque ofrece generalmente buenos resultados, la medida de distancia se definiría como:

$$\omega' = \omega \cdot (1 + u) \qquad [3.36]$$

La transformación lineal tiene dos efectos no deseados: el primero, que puede obtener pesos negativos; y el segundo, que ofrece valores no acotados, y por tanto es probable que los pesos puedan tomar valores muy extremos.

Por otra parte, existe el método exponencial que se basa en la medida de distancia:

$$\omega_k \cdot log\left(\frac{w_k}{d_k}\right) - w_k + d_k \qquad [3.37]$$

En este caso, siempre se generan pesos positivos, pero en cambio, puede haber una mayor distorsión de pesos nuevos respecto a los originales. Para este método, como para el siguiente *logit*, es recomendable establecer cotas a la transformación de los pesos originales, es decir, se buscan dos valores L y U tal que $L < h_k < U, k = 1, 2....n$ donde $h_k = w'_k/w_k$.

Los métodos *logit* lineal y truncado, también son métodos acotados, lo que significa que ofrecen límites superior e inferior sobre las relaciones de peso $\left(\frac{w_k}{d_k}\right)$, que se denominan pesos g y que permiten controlar las transformaciones de pesos extremos. Estas relaciones no son arbitrarias y dependen de las variables de calibración elegidas, por lo general, para determinar las cotas se realiza un proceso iterativo de descubrimiento de los límites. Como propone Rao (1996), se establece inicialmente un intervalo amplio para h_k que se va reduciendo progresivamente mientras se consiga una solución.

El procedimiento es fácilmente generalizable a múltiples dimensiones. Así, si suponemos 3 dimensiones para el caso lineal, por ejemplo, provincia, grupo de sexo y edad, los nuevos pesos se calcularían según:

$$\omega'_k = \omega_k \cdot (1 + x_{prov} + y_{sexo} + z_{edad}) \qquad [3.38]$$

donde las cantidades x_{prov}, y_{sexo} y z_{edad} serían las incógnitas a resolver, que aunque pueden tomar cualquier signo, normalmente al sumarse ofrecen una magnitud cercana a cero, para así satisfacer mejor la condición de variación mínima en la transformación de los pesos.

Elección de datos auxiliares para la calibración

La elección de las fuentes de información para la calibración no es sencilla, se asume que tanto las variables y sus totales son completos y exactos, aunque en la práctica no suele cumplirse. En los casos más extremos, los errores o distorsiones en la información auxiliar pueden llegar a dañar seriamente los pesos calibrados.

Los métodos de calibración permiten mejorar la estimación de parámetros mediante la incorporación de fuentes adicionales[45], mitigan los problemas de falta de respuesta, mejoran el ajuste de los pesos poblacionales y permiten

[45] Por ejemplo, se pueden tomar los totales de calibración de los resultados de otra encuesta, como hace la Oficina Central de Estadísticas de Irlanda para las estimaciones de su encuesta de población activa, que usa para calibrar los datos de las estadísticas de Eurotat de ingresos y condiciones de vida (EU-SILC). Esta última fuente incorpora, entre otros, nivel de ingresos, tasa de pobreza, inclusión social o pensiones.

obtener estimaciones consistentes entre las variables auxiliares y la muestra. Pero para que el proceso sea válido, es necesario asegurar la consistencia y calidad de la información de las variables auxiliares, por tanto debe asegurase que:

- Las estimaciones usadas como variables auxiliares sean (casi) imparciales y provengan de una muestra que sea, como mínimo, del mismo tamaño.
- Que ambas encuestas sean consistentes en cuanto a las magnitudes que miden, por ejemplo si ambas manejan ingresos, estas debe referirse a la misma magnitud.

Evaluación de la calidad del proceso

El control de la validez de todos pasos del proceso desde control de calidad y coherencia de las fuentes, pasando por las técnicas de modelado y terminando por los algoritmos de imputación, tienen un enorme impacto en la calidad de los resultados. Aún cuando se aseguren ciertos niveles requisitos de coherencia en los datos de entrada, los resultados se deben validar en función de su capacidad de proporcionar estimaciones fiables y precisas.

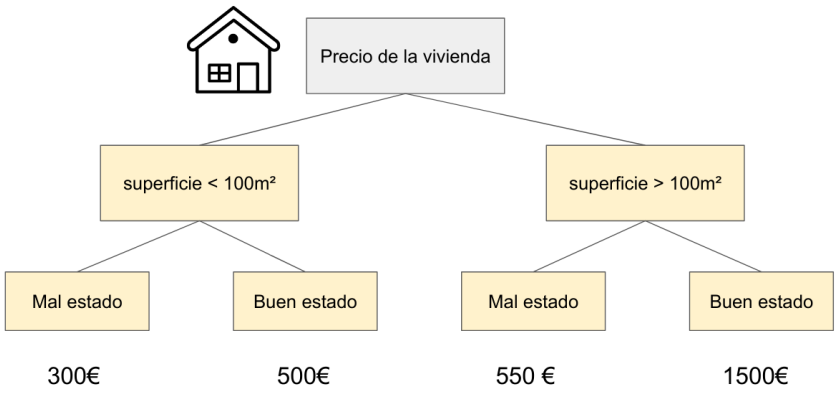
Rässler (2012) propone un marco de trabajo para la evaluación de la calidad a través de cuatro niveles de validación: 1) las distribuciones marginales y conjuntas de las variables del fichero donante se preservan en el fichero final; 2) la estructura de la correlación y los momentos altos de las variables se preservan después de la correspondencia estadística; 3) la distribución verdadera conjunta de todas las variables se refleja en el fichero final; y 4) los valores verdaderos pero desconocidos de la variable Z del fichero receptor se reproducen.

En el proceso es esencial tener en cuenta la incertidumbre, y en particular aquella asociada a las asunciones *a priori* implícitas en el modelo. Ante las limitaciones metodológicas, se proponen dos enfoques (D'Orazio *et al.*, 2006): 1) estimar la incertidumbre en las estimaciones finales, lo que está generalmente enfocado a macro-objetivos (estimación de coeficientes de correlación y tablas de contingencia); 2) Centrado en la identificación de información auxiliar que permita reducir la incertidumbre y pueda relajar las condiciones de independencia condicional.

Anexo 3b. Algoritmo Random Forests

Los modelos de tipo Random Forests, o bosques aleatorios, son un tipo de modelo de aprendizaje estadístico que usan árboles de decisión o regresión. Para la valoración de la vivienda, un modelo de regresión basado en árboles, estima el precio en base a una serie de "reglas de decisión". En el ejemplo gráfico de la Figura 3.28 se muestra cómo un modelo de tipo árbol estima el precio de una vivienda, para llegar al precio final se siguen una serie de reglas que acotan el precio en base a sus características.

Figura 3.28. odelo de valoración basado en un árbol de decisión simple



Fuente: elaboración propia.

A partir de los datos de entrada, las reglas de decisión (cortes) del árbol se calculan en función de su capacidad para reducir de la entropía en los grupos formados después del corte. Este desorden se puede expresar como impureza de los subárboles generados, ganancia/pérdida de información, coeficiente Gini o la varianza. Los modelos de árbol de regresión son una alternativa eficaz al análisis de regresión múltiple (Breiman, 2017; Fan *et al.*, 2006).

En conjuntos grandes y con una gran cantidad de variables, los modelos de árboles simples pueden incurrir en problemas de infrajuste o sobreajuste. Para resolver estos problemas, se aplican los enfoques basados en ensamblados de árboles como los basados en *bagging*[46], *boosting*[47] o *stacking*[48].

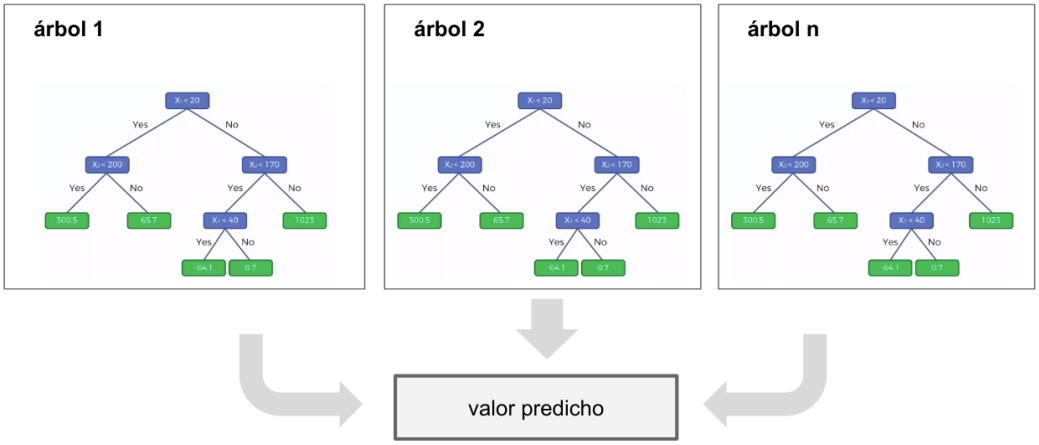
La técnica de Random Forests (Breiman, 2001), es un modelo de árbol de tipo ensamblado y fue desarrollada originalmente por Leo Breiman y Adele Cutler. Combina la idea de *bagging* y la selección aleatoria de atributos para construir

[46] El *bagging* consiste en combinar modelos distintos en paralelo con el objetivo de reducir la varianza.
[47] El *boosting* construye una secuencia de modelos predictivos orientados a corregir los errores del modelo anterior, para mejorar la precisión general del modelo final.
[48] El *stacking* utiliza las predicciones de varios modelos base para entrenar un modelo objetivo que tiene un mejor rendimiento predictivo que los originales.

una colección de árboles de decisión. Este algoritmo puede utilizarse para estimar valores categóricos (binarios o multiclase), denominado árbol de clasificación, o también para estimar una magnitud continua, dónde se lo denomina como árbol de regresión.

El modelo de *bagging* funciona construyendo una multitud de árboles de decisión en el momento del entrenamiento y generando una predicción media (regresión) de los árboles individuales (Figura 3.29). Este enfoque resuelve la tendencia al sobreajuste de otras modalidades de árboles ensamblados[49].

Figura 3.29. Esquema general de un modelo basado en Random Forests



Fuente: elaboración propia.

La idea esencial del *bagging* es promediar muchos modelos ruidosos pero aproximadamente insesgados (Hastie *et al.*, 2017), para producir un modelo combinado capaz de reducir la varianza. Si bien este proceso no fuerza el uso de un tipo de modelo en concreto, los modelos de árbol son los candidatos ideales para el *bagging*, dado que pueden registrar estructuras de interacción complejas en los datos, y si crecen con suficiente profundidad, tienen relativamente bajo sesgo. Dado que los árboles son notoriamente ruidosos, se benefician enormemente de la estimación basada en el promedio.

Cada árbol se construye a través de los siguientes pasos:

- Sea N el número de casos de prueba, y M es el número de variables en el clasificador.

[49] El sobreajuste se refiere la incapacidad de generalizar de un modelo, que no obstante muestra bajos niveles de error sobre el conjunto de datos de entrenamiento. Por tanto el modelo se ha especializado en replicar el dato de entrenamiento no en crear reglas que permitan evaluar correctamente otros datos.

- Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado, m debe ser un número mucho menor que M.
- Se elige un conjunto de entrenamiento para este árbol y se usa el resto de los casos de prueba para la estimación del error.
- Para cada nodo del árbol, se eligen aleatoriamente m variables, en las cuales basar la decisión y se calculan las mejores particiones del conjunto de entrenamiento, a partir de las m variables.

Finalmente, la inferencia se realiza recorriendo descendente cada árbol de decisión, tomando el valor del nodo terminal al que se llega a partir de las variables de entrada. Este proceso se realiza sobre todos los árboles en el ensamblado, y la estimación final se calcula como el promedio de las estimaciones individuales.

Antipov y Pokryshevskaya (2012), en su análisis sobre el modelo de *Random Forests* aplicado al ámbito inmobiliario, argumenta que esta técnica es una de las más adecuadas para la valoración de la vivienda por varios motivos:

- Muestra buenos resultados comparado con otras técnicas, como las máquinas de soporte vectorial (SVM), redes neuronales u otro tipo de modelos de árboles complejos como el *boosting*.
- Maneja satisfactoriamente variables categóricas con un gran número de niveles, sin incrementar el número de parámetros (como sería el caso de las redes neuronales que requieren la creación de variables ficticias *dummies*) lo que reduce la posibilidad del sobreajuste al introducir un gran volumen de variables dicotómicas.
- Funciona correctamente cuando existen valores ausentes, y no requiere procesos de imputación ni la eliminación de observaciones por este motivo.
- El proceso de *bagging* hace que el modelo sea robusto ante valores atípicos, ya que aparecerán menos en el muestreo de creación de árboles individuales (muestras de *bootstrap*), y por tanto su influencia en los resultados se ve reducida.
- Al contrario de los modelos de árboles de regresión simple (como CART), la estimación es un único valor, no una serie de valores discretos derivados de una serie de reglas.
- Los árboles permiten la gestión de las no linealidades, la heterocedasticidad y los comportamientos diferenciados las variables para los distintos segmentos, que los modelos lineales multivariantes.
- No se requiere una especificación detallada *a priori*.
- Las predicciones se encuentran en los mismos rangos que las observadas, lo que reduce la posibilidad de sobrestimación de las viviendas.

- Es posible medir la importancia de los factores, a través de su capacidad de reducción marginal de los errores, por parte de cada variable explicatoria.

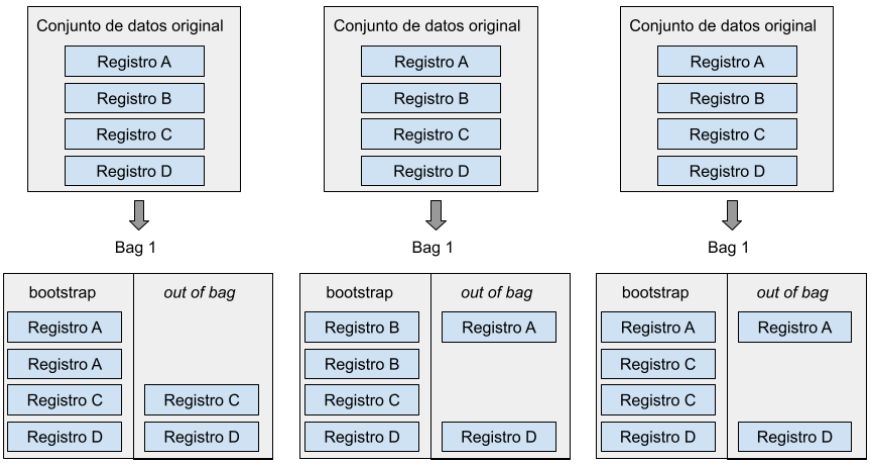
Muestra out-of-bag (OOB)

Una de las ventajas del método *Random Forests* es que puede generar métricas de error y de ajuste sin necesidad de un proceso de remuestreo previo que requiera dividir la muestra entre entrenamiento y validación. Para cada árbol construido, el algoritmo divide la muestra en dos conjuntos disjuntos: la *in bag* que se utiliza para construir el árbol de decisión y la *out of bag*, que se utiliza para calcular métricas sobre el modelo.

El error OOB, también es un método para medir el error de la predicción de *Random Forests*, aunque también es posible aplicarlo a otros tipos árboles de decisión basados *boosting*, y otros algoritmos de aprendizaje automático que utilizan agregación de tipo *bootstrap*[50](Hastie *et al.*, 2017).

Para comprender mejor la cuestión, en la Figura 3.30 se describe como se crea el conjunto OOB en el proceso de bagging. Primero se divide el conjunto en registros para entrenar (*in bag*) y para evaluar (*out of bag*) que no se utilizan en el entrenamiento, pero si para el cálculo de métricas de ajuste y error. Se realiza un proceso de sobremuestreo para completar el número de instancias de la muestra *"in bag"*. Este proceso se repite para cada árbol creado, de manera que las métricas del modelo se calculan promediando las medidas individuales.

Figura 3.30. Proceso de bagging mediante muestreo con reemplazo



Fuente: elaboración propia.

Con un suficiente número de árboles, las métricas OOB y las producidas la técnica

[50]Los métodos *bootstrap*,consisten en crear múltiples muestras de entrenamiento aleatorias con reemplazo de un conjunto de datos, para luego construir modelos a partir de estas muestras y combinar sus predicciones para mejorar la estabilidad y la precisión del modelo final.

de remuestreo denominadas de validación cruzada (LeCun *et al.*, 2015) producen una estimación similar.

Anexo 3c. Modelos GAM

Los modelos lineales tradicionales son simples, pero habitualmente adolecen de problemas cuando se aplican a situaciones reales, puesto que, los efectos rara vez suelen ser lineales (Hastie *et al.*, 2017).

La relación no lineal entre los predictores y la variable objetivo se puede controlar a través de flexibilizar los coeficientes de regresión, haciéndolos que sean función de las covariables, en lugar de constantes (Hastie y Tibshirani, 2017). Las funciones sobre las que se construyen estos coeficientes, de tipo funcional, se denominan funciones base, y son elemento central de los modelos aditivos generalizados.

En una regresión, un modelo aditivo generalizado tendría la forma siguiente:

$$E(Y|X_1, X_2, ..., X_p) = \alpha + f_1(X_1) + f_1(X_2) + ... + f_1(X_p) \quad [3.39]$$

donde $X_1, X_2, ..., X_p$ son los predictores, e Y la variable respuesta; las f_j son funciones de suavizado no paramétricas. Cada una de ellas se construye como una expansión de funciones base[51], para dar lugar a un modelo que se ajusta como una regresión simple de mínimos cuadrados.

En general, la medida condicional $\mu(X)$ de la variable de respuesta Y está relacionada con la función aditiva de los predictores mediante la función de enlace g:

$$g[\mu(X)] = \alpha + f_1(X_1) + f_1(X_2) + ... + f_1(X_p) \quad [3.40]$$

Ejemplo de función típicas son:

- $g(\mu)$ es el enlace identidad, se utiliza para mdoelos aditivos lineales con una respuesta gaussiana.
- $g(\mu) = logit(\mu)$ o $g(\mu) = probit(\mu)$ se utiliza para modelar probabilidades binomiales.
- $g(\mu) = log(\mu)$ se corresponde a modelos log-lineales o logarítmico-aditivos asociados a funciones de distribución de Poisson (conteos).

[51] La expansión se refiere a la suma de los resultados de las distintas funciones de suavizado

Los modelos aditivos se pueden entender como forma una extensión sobre los modelos lineales, que los hace más flexibles mientras que se mantiene su interpretabilidad. Sin embargo estos modelos pueden tener limitaciones en análisis con grandes volúmenes de datos y alta dimensionalidad (Hastie *et al.*, 2017), principalmente porque el método intenta estimar una función de suavizado para cada predictor. Se han planteado diversas aproximacion con el objeto de resolver los inconvenientes anteriores, como por ejemplo utilizar penalizaciones de tipo Lasso, denominadas COSSO (Lin y Zhang, 2006), o el método SpAM[52] (Ravikumar *et al.*, 2007).

Para grandes volúmenes de datos, también se puede aplicar métodos como el *boosting* para estimar la expansión de las funciones base. En los últimos años, se han publicado una serie de algoritmos que permiten estimar modelos GAM de forma eficiente en conjuntos grandes (Li y Wood, 2020; Wood *et al.*, 2015, 2017).

En el ámbito de la valoración de la vivienda existen diversas aplicaciones como el modelo desarrollado por Pace (1998) que intenta limitar el efecto de las no linealidades con un hedónico basado en GAM. A modo de ejemplo se puede señalar la aplicación al mercado de la vivienda en Alemania de Munger (2021), para Eslovenia encontramos a (Ulbl *et al.*, 2021), para Sudáfrica usando un método GAM jerárquico (Bax *et al.*, 2021).

[52] *Sparse Additive Models*.

Anexo 3d. Sumas calibración de la EPF

Las Tabla 3.17 muestra las sumas originales del conjunto de datos de la EPF, sobre las que se ha aplicado un proceso de suavizado exponencial para calcular las sumas de la calibración definitiva, y que se recogen en la Tabla 3.18.

Tabla 3.17. Totales originales para calibración EPF

Variable	2011	2012	2013	2014	2015	2016	2017	2018	2019
Total	428667	96%	95%	95%	99%	111%	113%	109%	122%
1 o 2 habitaciones	267959	106%	104%	106%	110%	120%	112%	118%	138%
3 habitaciones	126257	87%	78%	81%	84%	101%	115%	98%	99%
4 habitaciones	25964	89%	158%	82%	132%	144%	161%	131%	142%
5 o más habitaciones	8487	48%	39%	70%	39%	33%	40%	41%	68%
Menos de 60	174119	110%	93%	104%	108%	120%	105%	108%	122%
De 61 a 75	99464	77%	86%	97%	81%	89%	92%	93%	106%
De 76 a 90	86505	100%	93%	77%	117%	148%	155%	138%	174%
Más de 90	68579	98%	133%	95%	99%	109%	145%	119%	124%
Menos de 10 viviendas	81814	95%	114%	107%	99%	112%	130%	115%	113%
10 o más viviendas	346854	97%	90%	92%	99%	111%	109%	107%	124%
Chalé	13337	96%	173%	200%	210%	120%	158%	136%	154%
Casa media	362376	95%	92%	91%	102%	117%	120%	114%	127%
Casa económica	52954	104%	90%	86%	49%	68%	57%	65%	77%
Urbana alta	40882	97%	123%	71%	84%	84%	92%	61%	98%
Urbana media	365154	97%	95%	100%	102%	118%	119%	114%	123%
Urbana inferior	22631	81%	36%	46%	77%	51%	48%	110%	150%
Renta baja	188857	96%	92%	93%	87%	85%	79%	73%	54%
Renta media	186721	86%	95%	87%	110%	121%	127%	107%	130%
Renta alta	53089	117%	134%	149%	100%	186%	193%	213%	269%
Gasto bajo-medio	28825	112%	96%	163%	135%	187%	166%	128%	134%
Gasto medio-alto	170787	78%	67%	85%	105%	124%	128%	109%	112%
Gasto alto	229056	100%	104%	85%	90%	92%	98%	105%	123%
Ciudad de Madrid	264129	88%	83%	80%	88%	106%	98%	97%	103%
Resto CAM	164538	118%	127%	135%	129%	124%	152%	141%	172%
Zona densamente poblada	387080	95%	88%	86%	95%	107%	104%	100%	111%
Zona intermedia	26142	120%	172%	160%	128%	125%	169%	175%	213%
Zona diseminada	15446	97%	133%	192%	155%	196%	235%	215%	228%

Fuente: elaboración propia

Tabla 3.18. Totales suavizados exponencialmente para calibración EPF

Variable	2011	2012	2013	2014	2015	2016	2017	2018	2019
Total	428667	103%	102%	101%	100%	102%	109%	114%	114%
1 o 2 habitaciones	267959	105%	110%	112%	114%	117%	123%	124%	126%
3 habitaciones	126257	100%	93%	85%	83%	83%	92%	103%	100%
4 habitaciones	25964	105%	110%	116%	121%	126%	132%	137%	142%
5 o más habitaciones	8487	96%	68%	49%	55%	42%	33%	32%	32%
Menos de 60	174119	103%	106%	108%	110%	112%	115%	117%	118%
De 61 a 75	99464	101%	90%	89%	93%	88%	89%	91%	93%
De 76 a 90	86505	109%	116%	119%	117%	126%	141%	154%	159%
Más de 90	68579	103%	106%	109%	112%	115%	118%	121%	124%
Menos de 10 viviendas	81814	102%	103%	105%	107%	108%	110%	112%	114%
10 o más viviendas	346854	103%	103%	99%	98%	102%	109%	112%	112%
Chalé	13337	107%	108%	148%	181%	203%	168%	170%	160%
Casa media	362376	103%	103%	101%	99%	104%	114%	120%	121%
Casa económica	52954	97%	98%	91%	86%	64%	63%	57%	58%
Urbana alta	40882	100%	99%	102%	98%	96%	93%	91%	85%
Urbana media	365154	103%	103%	102%	104%	106%	114%	119%	120%
Urbana inferior	22631	106%	100%	73%	65%	77%	70%	64%	93%
Renta baja	188857	94%	89%	85%	83%	80%	77%	72%	67%
Renta media	186721	104%	105%	107%	107%	111%	117%	122%	123%
Renta alta	53089	121%	141%	160%	178%	175%	199%	218%	238%
Gasto bajo-medio	28825	104%	111%	111%	129%	135%	152%	160%	157%
Gasto medio-alto	170787	102%	91%	80%	84%	96%	112%	121%	117%
Gasto alto	229056	103%	104%	107%	99%	97%	97%	100%	105%
Ciudad de Madrid	264129	100%	95%	89%	85%	87%	97%	98%	98%
Resto CAM	164538	109%	118%	127%	136%	145%	154%	163%	172%
Zona densamente poblada	387080	101%	100%	95%	92%	95%	102%	105%	104%
Zona intermedia	26142	114%	128%	143%	157%	171%	185%	199%	213%
Zona diseminada	15446	116%	131%	147%	166%	181%	198%	216%	232%

Fuente: elaboración propia

Anexo 3e. Descriptivos de modelos hedónicos

Las Tablas 3.19 y 3.20 recogen los coeficientes de la regresión del mercado de la EPF, y las Tabla 3.23 los de regresión del modelo de oferta. Los coeficientes del modelo de conversión se muestran en la Tabla 3.25. En todos los casos anteriores, los términos de las funciones de suavizado de los modelos son altamente significativos, véanse las Tablas 3.22, 3.21 y 3.24 del presente anexo.

Tabla 3.19. Coeficientes regresión GAM - Modelo alquiler sobre EPF 2012 - 1/2

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	3.89	0.05	78.82	0.00	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	0.01	0.02	0.22	0.83	
TAMAMUMunicipio con 20.000 o más y menos de 50.000	-0.01	0.03	-0.25	0.80	
TAMAMUMunicipio con 10.000 o más y menos de 20.000	-0.09	0.03	-2.77	0.01	**
TAMAMUMunicipio con menos de 10.000 habitantes	-0.08	0.03	-2.56	0.01	*
TIPOEDIFVivienda unifamiliar adosada o pareada	-0.05	0.01	-5.14	0.00	***
TIPOEDIFCon menos de 10 viviendas	-0.07	0.02	-3.86	0.00	***
TIPOEDIFCon 10 ó más viviendas	-0.05	0.02	-3.00	0.00	**
TIPOEDIFOtros (destinado a otros fines o alojamien	-0.04	0.11	-0.38	0.70	
ZONARESUrbana alta	0.08	0.04	2.30	0.02	*
ZONARESUrbana media	0.02	0.03	0.43	0.66	
ZONARESUrbana inferior	0.02	0.04	0.50	0.62	
ZONARESRural industrial	-0.11	0.04	-2.76	0.01	**
ZONARESRural pesquera	-0.03	0.06	-0.56	0.57	
ZONARESRural agraria	-0.03	0.04	-0.84	0.40	
ANNOCONHace 25 ó más años	-0.01	0.01	-1.59	0.11	
DENSIZona intermedia	0.02	0.03	0.90	0.37	
DENSIZona diseminada	-0.05	0.03	-1.89	0.06	.
INTERINPSPDe 500 a menos de 1000 €	-0.01	0.02	-0.47	0.64	
INTERINPSPDe 1000 a menos de 1500 €	0.02	0.02	0.97	0.33	
INTERINPSPDe 1500 a menos de 2000 €	0.02	0.02	1.08	0.28	
INTERINPSPDe 2000 a menos de 2500 €	0.06	0.02	2.32	0.02	*
INTERINPSPDe 2500 a menos de 3000 €	0.03	0.03	1.26	0.21	
INTERINPSP3000 o más €	0.02	0.03	0.80	0.43	
NHABIT3 habitaciones	-0.03	0.02	-1.71	0.09	.
NHABIT4 habitaciones	-0.05	0.02	-2.67	0.01	**
NHABIT5 o más habitaciones	-0.05	0.02	-2.50	0.01	*
CAPROVNo	-0.10	0.02	-5.71	0.00	***
factorGASTOT6De.15.83.a.16.26	0.02	0.02	0.73	0.46	
factorGASTOT6De.16.26.a.16.62	0.04	0.02	1.75	0.08	.
factorGASTOT6De.16.62.a.17	0.02	0.02	0.84	0.40	
factorGASTOT6De.17.a.17.46	0.05	0.02	2.25	0.02	*
factorGASTOT6Más.de.17.46	0.09	0.02	3.75	0.00	***

Fuente: elaboración propia

Tabla 3.20. Coeficientes regresión GAM - Modelo alquiler sobre EPF 2012 - 2/2

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
CCAAAragón	0.09	0.03	2.94	0.00	**
CCAAAsturias, Principado de	0.17	0.04	4.76	0.00	***
CCAABalears, Illes	0.30	0.04	7.49	0.00	***
CCAACanarias	0.06	0.02	2.77	0.01	**
CCAACantabria	0.28	0.03	9.75	0.00	***
CCAACastilla y León	-0.01	0.02	-0.46	0.64	
CCAACastilla-La Mancha	0.05	0.02	2.52	0.01	*
CCAACataluña	0.29	0.02	17.87	0.00	***
CCAAComunitat Valenciana	-0.02	0.01	-1.24	0.22	
CCAAExtremadura	-0.10	0.03	-2.95	0.00	**
CCAAGalicia	0.04	0.02	1.95	0.05	.
CCAAMadrid, Comunidad de	0.22	0.02	11.59	0.00	***
CCAAMurcia, Región de	-0.05	0.02	-2.56	0.01	*
CCAANavarra, Comunidad Foral de	0.43	0.03	12.23	0.00	***
CCAAPaís Vasco	0.50	0.06	8.12	0.00	***
CCAARioja, La	0.23	0.09	2.64	0.01	**
CCAACeuta	0.36	0.09	3.95	0.00	***
CCAAMelilla	0.07	0.09	0.74	0.46	

Fuente: elaboración propia

Modelo de alquiler EPF para 2012

Tabla 3.21. Términos de la función de suavizado - Modelo alquiler sobre EPF 2012

Término	edf	Ref. df	F	p-value	signif.
s(SUPERF)	7.59	8.47	63.43	0	***

Fuente: elaboración propia

Modelo de idealista para 2012

Tabla 3.22. Términos de la función de suavizado - Modelo idealista 2012

Término	edf	Ref. df	F	p-value	signif.
s(SUPERF2)	8.74	8.98	384.83	0	***

Fuente: elaboración propia

Tabla 3.23. Coeficientes regresión GAM - Modelo oferta 2012

Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	4.44	0.06	68.49	0.00	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	0.07	0.01	5.66	0.00	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	0.03	0.01	2.18	0.03	*
TAMAMUMunicipio con 10.000 o más y menos de 20.000	0.04	0.02	2.25	0.02	*
TAMAMUMunicipio con menos de 10.000 habitantes	-0.15	0.02	-9.19	0.00	***
TIPOEDIFVivienda unifamiliar adosada o pareada	0.26	0.01	29.96	0.00	***
ZONARESUrbana alta	-0.24	0.02	-11.65	0.00	***
ZONARESUrbana media	-0.25	0.03	-8.50	0.00	***
ZONARESUrbana inferior	-0.22	0.04	-5.35	0.00	***
ANNOCONHace 25 ó más años	0.01	0.01	2.40	0.02	*
DENSIZona intermedia	-0.05	0.01	-3.97	0.00	***
DENSIZona diseminada	-0.11	0.02	-7.22	0.00	***
INTERINPSPDe 500 a menos de 1000 €	0.03	0.01	1.73	0.08	.
INTERINPSPDe 1000 a menos de 1500 €	0.11	0.02	7.21	0.00	***
INTERINPSPDe 1500 a menos de 2000 €	0.25	0.02	14.54	0.00	***
INTERINPSPDe 2000 a menos de 2500 €	0.38	0.02	18.43	0.00	***
INTERINPSPDe 2500 a menos de 3000 €	0.46	0.03	17.05	0.00	***
INTERINPSP3000 o más €	0.47	0.03	15.01	0.00	***
NHABIT3 habitaciones	-0.05	0.01	-4.61	0.00	***
NHABIT4 habitaciones	-0.03	0.01	-2.63	0.01	**
NHABIT5 o más habitaciones	0.01	0.01	0.58	0.56	
CAPROVNo	-0.22	0.01	-17.73	0.00	***
factorGASTOT6De.16.26.a.16.62	-0.05	0.06	-0.72	0.47	
factorGASTOT6De.16.62.a.17	-0.04	0.06	-0.64	0.52	
factorGASTOT6De.17.a.17.46	-0.06	0.06	-1.00	0.32	
factorGASTOT6Más.de.17.46	-0.02	0.06	-0.39	0.69	

Fuente: elaboración propia

Modelo de correspondencia oferta-alquiler unifamiliar 2012

Tabla 3.24. Términos de la función de suavizado - Modelo conversion 2012

Término	edf	Ref. df	F	p-value	signif.
s(preciom2_anualpred)	8.99	9	1641.03	0	***
s(SUPERF2)	9.00	9	25917.82	0	***

Fuente: elaboración propia

Tabla 3.25. Coeficientes regresión GAM - Modelo conversion 2012

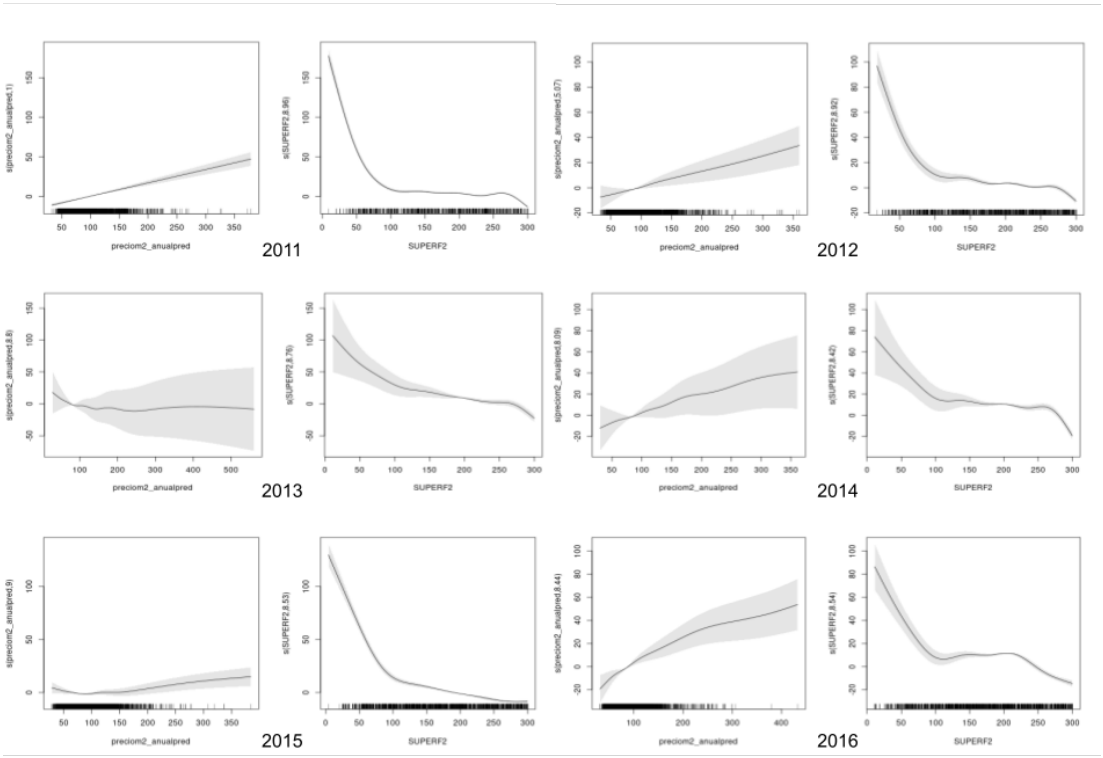
Coeficiente	Estimate	Std.Err	t value	p-value	signif.
INTERCEPT	85.78	0.55	156.76	0	***
TAMAMUMunicipio con 50.000 o más y menos 100.000 h	6.97	0.09	77.16	0	***
TAMAMUMunicipio con 20.000 o más y menos de 50.000	8.23	0.14	59.88	0	***
TAMAMUMunicipio con 10.000 o más y menos de 20.000	7.18	0.19	37.44	0	***
TAMAMUMunicipio con menos de 10.000 habitantes	7.20	0.25	28.71	0	***
TIPOEDIFCon 10 ó más viviendas	1.65	0.07	23.46	0	***
TIPOCASACasa económica o alojamiento	-4.88	0.12	-40.46	0	***
ZONARESUrbana alta	-1.66	0.32	-5.14	0	***
ZONARESUrbana media	-1.71	0.38	-4.47	0	***
ZONARESUrbana inferior	-7.62	0.41	-18.46	0	***
ANNOCONHace 25 ó más años	0.84	0.05	15.79	0	***
DENSIZona intermedia	-6.42	0.14	-45.88	0	***
DENSIZona diseminada	-16.65	0.26	-63.47	0	***
INTERINPSPDe 500 a menos de 1000 €	0.65	0.12	5.44	0	***
INTERINPSPDe 1000 a menos de 1500 €	3.41	0.13	26.00	0	***
INTERINPSPDe 1500 a menos de 2000 €	7.97	0.17	47.63	0	***
INTERINPSPDe 2000 a menos de 2500 €	12.57	0.19	65.14	0	***
INTERINPSPDe 2500 a menos de 3000 €	11.46	0.29	38.98	0	***
INTERINPSP3000 o más €	8.30	0.35	23.56	0	***
NHABIT3 habitaciones	-0.95	0.06	-15.42	0	***
NHABIT4 habitaciones	0.65	0.12	5.39	0	***
NHABIT5 o más habitaciones	1.28	0.29	4.49	0	***
CAPROVNo	-11.86	0.09	-138.88	0	***
factorGASTOT6De.16.26.a.16.62	8.03	0.46	17.44	0	***
factorGASTOT6De.16.62.a.17	11.33	0.39	29.40	0	***
factorGASTOT6De.17.a.17.46	10.73	0.38	28.60	0	***
factorGASTOT6Más.de.17.46	14.08	0.38	37.26	0	***

Fuente: elaboración propia

Funciones de suavizado en modelo de correspondencia

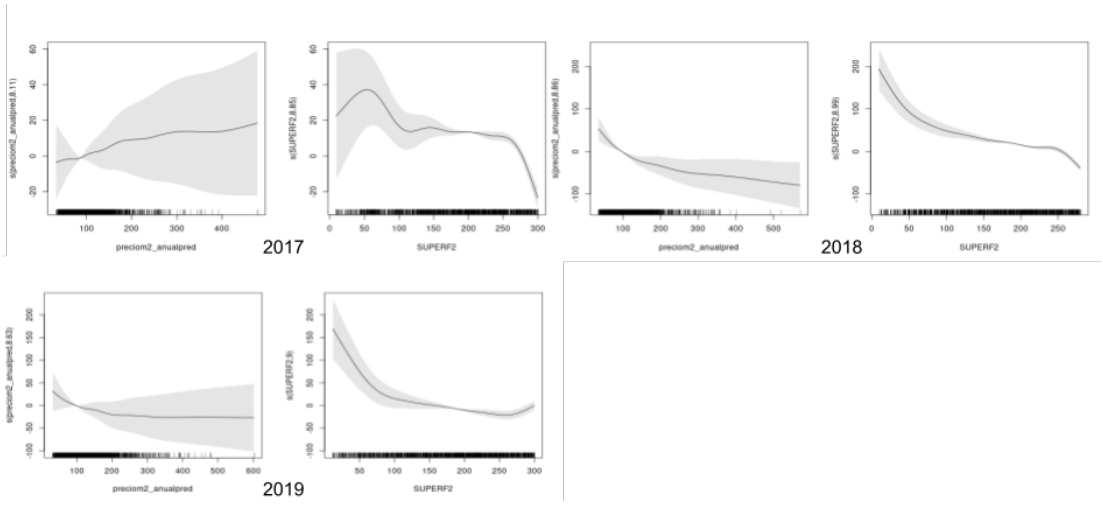
Las Figuras 3.31 y 3.32 muestran el comportamiento de las funciones de suavizado (*s*) de los modelos GAM de correspondencia, sobre el precio de oferta (preciom2_anualpred) y la superficie útil (SUPERF2).

Figura 3.31. Funciones de suavizado (s), vivienda unifamiliar: 2011-2016



Fuente: elaboración propia.

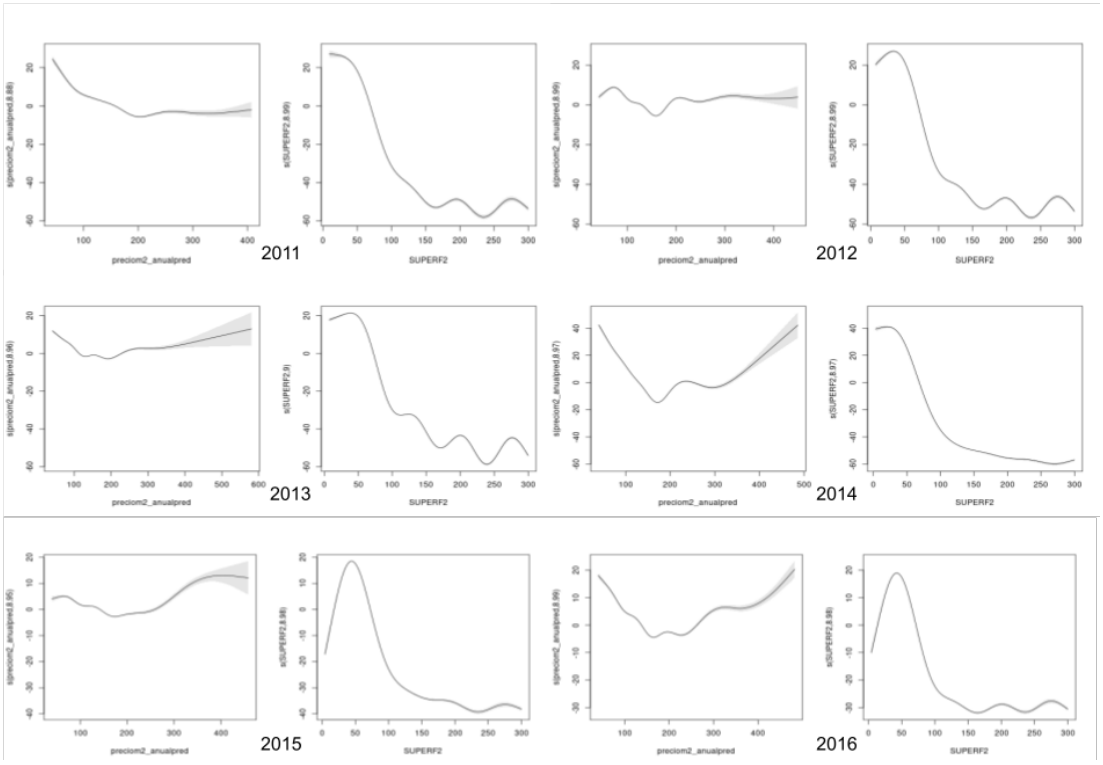
Figura 3.32. Funciones de suavizado (s), vivienda unifamiliar: 2017-2019



Fuente: elaboración propia.

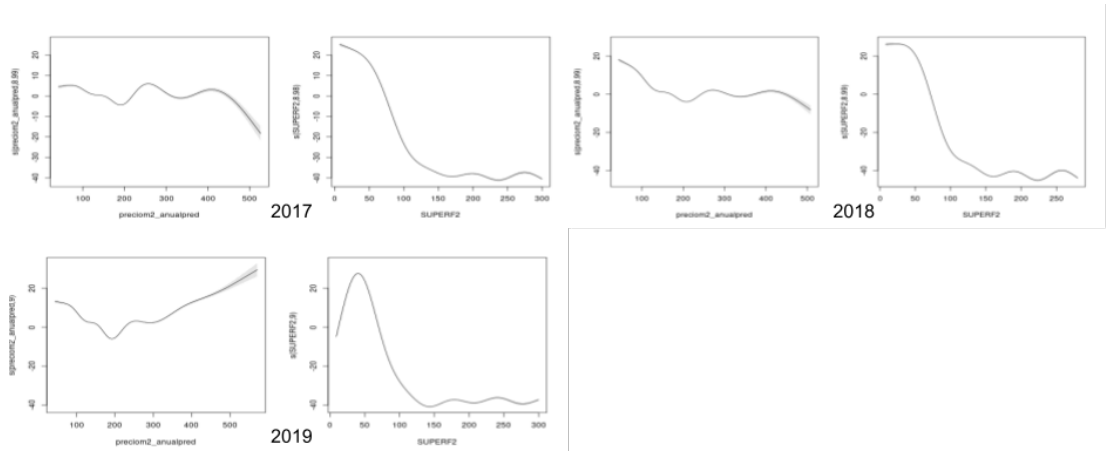
Las Figuras 3.33 y 3.34 muestran el comportamiento de las funciones de suavizado para las viviendas plurifamiliares. Se puede comprobar que las funciones de suavizado son mucho más significativas que en el caso de la vivienda plurifamiliar.

Figura 3.33. Funciones de suavizado (s), vivienda plurifamiliar: 2011-2016



Fuente: elaboración propia.

Figura 3.34. Funciones de suavizado (s), vivienda plurifamiliar:2017-2019



Fuente: elaboración propia.