Capítulo 5

Modelo hedónico de oferta

"Las cosas valen lo que se pagan por ellas en una venta."

— Edward Coke, jurista británico

5.1 Introducción

En los capítulos anteriores se ha trabajado en un modelo que represente la relación que guardan los precios y la estructura de pesos muestrales para los colectivos del alquiler y los de oferta (capítulos 2 y 3). La limitación de estos modelos procede de que varias de las fuentes de información sobre las que se trabaja, como el censo o la EPF, no cuentan con el suficiente desglose funcional, geográfico y temporal.

No obstante, dado que ya se conoce la relación funcional general entre poblaciones y precios, y se dispone de registros de oferta desagregados, se pueden mitigar las limitaciones anteriores mediante un modelo hedónico de oferta altamente detallado. Este modelo permitirá trasladar este nivel de desglose a los precios de mercado, calculados según la metodología de los capítulos anteriores.

El modelo de oferta a construir persigue un alto de nivel de fiabilidad, capacidad de ajuste y desglose, para todos los distintos estratos que componen la población y supone una verdadera innovación en el área. Por tanto, los modelos de aprendizaje estadístico son los candidatos ideales a aplicar en la metodología, en particular los modelos de árboles ensamblados de tipo Random Forests, a tenor de los resultados de distintos estudios que demuestran su superioridad para este caso de uso, frente otras técnicas de modelización (Alfaro Navarro *et al.*, 2020; Antipov y Pokryshevskaya, 2012; Graczyk *et al.*, 2010; Truong *et al.*, 2020).

Los modelos de valoración basados en datos de oferta de portales inmobiliarios se han popularizado en las últimas dos décadas, con una numerosa literatura

de casos en distintas geografías. Por ejemplo, en España (Alfaro Navarro *et al.*, 2020; Baldominos *et al.*, 2018; Del Cacho, 2010; Larraz y Poblacion, 2013); en Estambul (Turquía) (Özsoy y Şahin, 2009); en Montreal (Canadá) (Pow *et al.*, 2014); la República Checa (Larraz y Poblacion, 2013); en China (Truong *et al.*, 2020); (Clark y Lomax, 2018) en Reino Unido y (Pérez-Rave *et al.*, 2019) en Colombia.

Una de las primeras referencias en aplicar *Random Forests* en la valoración de vivienda es la de Antipov (Antipov y Pokryshevskaya, 2012), que argumenta su idoneidad comparado con otros métodos. En su estudio trabaja sobre un conjunto de datos de San Petersburgo (Rusia) y compara su rendimiento con 10 algoritmos de aprendizaje automático, evaluando su comportamiento general y en distintos segmentos de la muestra, adicionalmente, propone otras mejoras basadas en la aplicación de coeficientes de corrección en los distintos segmentos de la población. Čeh *et al.* (2018) comparan el desempeño de un modelo hedónico basado en regresión múltiple para viviendas en Lubjiana (Eslovenia), indicando que esta aproximación logra mejores resultados en todas las métricas habituales de evaluación de ajuste y error. Es interesante la construcción de una serie de componentes de accesibilidad, a través de PCA (Pearson, 1901), como variables auxiliares. Por su parte, Hong (2020) aplica *Random Forests* sobre viviendas del barrio de Gangnam en Seúl (Corea del Sur). En su estudio, el modelo de árboles reduce a una cuarta parte el error absoluto del modelo hedónico de regresión. En Noruega, Hjort (2022) estudian las mejoras de precisión al usar modelos de ensamblados de árboles, en este caso comparando el error en la valoración de carteras masivas de inmuebles (AVM).

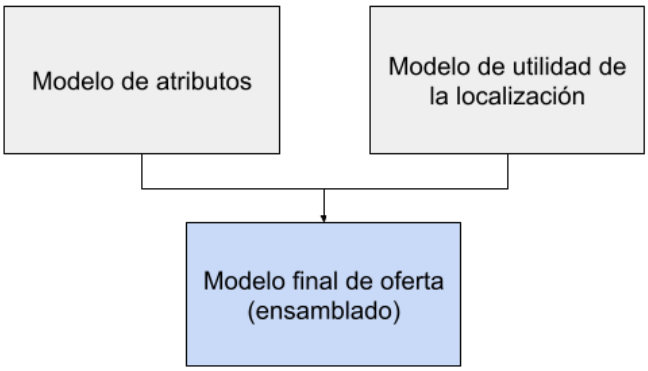
Para el caso español, se han publicado recientemente tres estudios donde se han usado modelos de esta naturaleza. Rico y Taltavull (2021) construyen un modelo de bosques aleatorios sobre un inmuebles de la provincia de Alicante, centrado en aspecto de la explicabilidad; Baldominos (2018) utiliza diferentes técnicas de aprendizaje automático, entra ellas Random Forests, para construir un modelo de regresión de precios de la vivienda para Madrid; Alfaro et al. (2020) construyen un modelo de valoración para 433 municipios, probando distintos modelos ensamblados, y logrando el mejor desempeño de todos con Random Forests.

Sin embargo, estos modelos requieren un establecimiento adecuado de los hiperparámetros del modelo (Antipov y Pokryshevskaya, 2012). Además, debe tenerse en cuenta que la precisión del modelo es sensible a los parámetros establecidos por el analista, según sugieren Prinzie y Van den Poel (2008), que centran su investigación sobre modelos de árboles. Otra cuestión relevante

sobre los árboles de regresión es su todavía limitada adopción en el campo econométrico, principalmente por su menor interpretabilidad que los métodos tradicionales. Si bien, en los últimos años ha habido avances importantes en este sentido (Lundberg *et al.*, 2018; Lundberg y Lee, 2017), con la introducción de técnicas que permiten explicar el comportamiento de los modelos de aprendizaje estadístico.

Como se presentaba en los capítulos 3 y 5, los inductores de precio de la vivienda son múltiples. Hill (2013) los reduce a los dos más relevantes: características y la localización. La metodología seguida para el modelo de oferta propone construir un modelo para cada inductor, junto con un tercer modelo ensamblado que los une, de forma que, se puedan incorporar tanto las contribuciones de los dos aspectos como las de sus interacciones. La Figura 5.1 representa gráficamente el proceso.

Figura 5.1. Diagrama general del modelo hedónico de oferta



Fuente: elaboración propia.

El capítulo se estructura en dos partes: primeramente, se describe el planteamiento metodológico para construir el modelo de oferta, centrado en cual es la mejor forma de elaborarlo, si sobre un modelo único o combinando varios; en segundo lugar, se presentan los resultados obtenidos, centrando la discusión en el aporte de la combinación de modelos, el ajuste espacial, el control de sesgos y la interpretabilidad.

5.2 Metodología

El proceso de creación de modelos hedónicos de precios de la vivienda presenta varias dificultades: la inexistencia de una forma funcional canónica, el comportamiento de mercado de los distintos segmentos de producto (unifamiliar, plurifamiliar) y la influencia de la dimensión espacio-temporal. Para intentar incorporar, en la mayor medida posible, el efecto de estos aspectos, se han construido diferentes modelos hedónicos por cada tipo de inmueble que se "ensamblan" para obtener la mejor estimación.

El concepto de ensamblado, o consenso de modelos, se fundamenta sobre la capacidad de construir un estimador fuerte mediante la agregación de una serie de modelos débiles o base. Estos últimos se pueden construir mediante diversas técnicas (árboles de decisión, redes neuronales, regresiones simples, etc.), para el modelo denominado de ensamblado combine todos los resultados que aproveche los distintos aspectos positivos de los modelos individuales, y compense sus errores (Zhou, 2021). La idea inicial de ensamblado es bastante antigua, y una de sus primeras menciones se puede encontrar en la investigación de Hansen y Salamon (1990), que demostró que la predicción combinada de una serie de modelos mejoraba los resultados de un clasificador individual.

En el campo del aprendizaje estadístico, los ensamblados comenzaron a popularizarse a partir de la década de los 2000, y se han aplicado eficazmente a varios campos como el sector farmacéutico, la banca, los sistemas de recomendación de contenidos o el control del fraude (Seni y Elder, 2010).

La investigación de Schapire (1990) prueba que la precisión de los modelos débiles puede multiplicarse a través de modelos agregados, denominados fuertes. Este trabajo da lugar a la familia de modelos denominados de *boosting* adaptativos, como por ejemplo Adaboost (Freund *et al.*, 1999), que ha sido uno de la más influyentes dentro de esta categoría.

Aunque existen trabajos teóricos sobre los métodos más comunes: *stacking*, *boosting* y *bagging*, no existe un entendimiento completo de los mecanismos subyacentes de los métodos, aunque los resultados empíricos indican que estas aproximaciones no adolecen de problemas de sobreajuste (Zhou, 2021) después de un número suficientemente grande de iteraciones, y en ocasiones, son capaces de reducir el error. Al ser métodos no paramétricos, el rendimiento de estos modelos se realiza mediante el estudio de la descomposición de sesgo-varianza. Para los algoritmos basados en *bagging*, se conoce que son eficaces en la reducción de la varianza, por tanto, ideales para aplicarlo en conjuntos con una alta varianza (este es uno de los motivos de usar Random Forests de forma

extensiva en la presente investigación). Adicionalmente, los modelos de *boosting* son capaces de reducir ambos factores de una manera eficaz (Hastie *et al.*, 2017).

En nuestro caso, se crea un modelo de ensamblado a medida con el objeto de controlar adecuadamente los fenómenos de la varianza y el sesgo (Zhou, 2021), bajo un principio de simplicidad, puesto que, se ha observado que una mayor complejidad en el ensamblado está asociada a resultados más pobres (Graczyk *et al.*, 2010). El método combina dos modelos para reducir de forma secuencial la varianza, controlando el sesgo introducido en cada paso.

Se unen dos modelos anuales, uno basado en atributos y otro basado en la localización, que hacen un total de 27 modelos para la serie histórica, 3 por cada uno de los 9 años de la serie[1]. Al especializar los dos procesos, se intenta evitar que un modelo único pase por alto alguna de las variables relevantes de la muestra, maximizando el aprovechamiento de las contribuciones marginales de las características en los dos ámbitos más importantes: los atributos constructivos y el área en la que se encuentra la vivienda. Desde un punto de vista inmobiliario-urbanístico, esta división se refiere al desglose de los dos atributos fundamentales que forman el precio de la vivienda, el precio del suelo y el precio de la construcción (o vuelo).

Dada la complejidad del conjunto de datos, en términos de atributos y variedad de submercados de la vivienda, se decide utilizar la técnica de modelado de regresión basado en árboles. Se ha optado por el algoritmo *Random Forests*[2] (Breiman, 2001), por su capacidad de gestionar no linealidades, precisión, velocidad de convergencia y menor tendencia al sobreajuste.

Una cuestión importante a tener en cuenta en la evaluación de resultados, es que los modelos de *bagging* no producen modelos sesgados, en el sentido estricto de que la media de los errores tiende a ser cero, pero si comportan otros tipos de sesgos. A este respecto, Breiman (1996) afirma que *Random Forests* reduce la varianza pero no actúa de forma eficaz sobre el sesgo del modelo, puesto que, tiende a modelar correctamente los valores medios (representados por las hojas finales del árbol), pero tiene dificultad para predecir los casos extremos. Se profundizará en esta cuestión en el epígrafe 6.3.2 del próximo capítulo, dedicado a los sesgos en el modelo hedónico final.

[1] Los tres modelos se corresponden al de atributos, localización y ensamblado.

[2] Para este caso se ha utilizado la versión denominada ranger (Wright y Ziegler, 2015), por su capacidad de trabajar eficientemente con grandes volúmenes de datos con alta dimensionalidad.

5.2.1 Modelo hedónico de características

El modelo hedónico de características construye la relación entre el precio y los atributos de la vivienda. Como se muestra en la Tabla 5.1 y en la Tabla 5.2, los 43 atributos atienden a características físicas del inmueble y de la finca, dinámicas de mercado de la zona, calidad y estado de conservación del inmueble.

Tabla 5.1. Variables modelo de atributos

Categoría	Variable	Fuente	Descripción	Inmueble
Area	CLUSTER	calculado	Tipo de zona siguiendo una clasificación propia	Unifamiliar
Calidad	CADASTRALQUALITYID	catastro	Calidad de la construcción	Ambos
Edificio	AMENITYID	idealista	Tipo de instalaciones de la finca	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
	CONSTRUCTIONYEAR	catastro	Año de construcción de la propiedad	Ambos
	DWELLING_COUNT	catastro	Número de inmuebles en la finca	Plurifamiliar
	FLOOR_POSITION	idealista	Posición del piso dentro del edificio	Plurifamiliar
	HASDOORMAN	idealista	Tiene portero	Unifamiliar
	HASGARDEN	idealista	Tiene jardín	Unifamiliar
	HASLIFT	idealista	Tiene ascensor	Plurifamiliar
	HASSWIMMINGPOOL	idealista	¿Tiene su edificio una piscina?	Ambos
	MAXBUILDINGFLOOR	idealista	Número de pisos en edificio	Ambos
Fecha	PERIOD	idealista	Código mes año en formato YYYYMM	Ambos
Mercado	CHANNELID	idealista	Canal de comercialización (agencia / particular)	Ambos
	LEADS_RESIDENTIAL	idealista	Número de contactos medios en la zona	Unifamiliar
	ONMARKET_RENT	idealista	Número de inmuebles en alquiler en la zona	Unifamiliar
	ONMARKET_SALE	idealista	Número de inmuebles en venta en la zona	Unifamiliar
	RENTSALE_RATIO	idealista	Proporción de inmuebles en alquiler / venta	Unifamiliar

Fuente: elaboración propia

Para evitar, en buena medida, el sesgo ocasionado por no incluir la información de la zona (precio del suelo), se incluye la variable denominada *Cluster*, que indica la categoría de zona a la que pertenece la seccion censal en la que se encuentra el

anuncio. Estas categorías se estiman de forma automática mediante un algoritmo de análisis *cluster*, descrito en el Anexo 5a de este capítulo.

Otra aportación original de este modelo de atributos es la incorporación de las características de competencia del mercado. Existen evidencias empíricas, como en el modelo propuesto por Fuss y Koller (2016), que señalan la capacidad predictiva de las variables que describen las dinámicas de los mercados locales.

Tabla 5.2. Variables modelo de atributos (continuación)

Categoría	Variable	Fuente	Descripción	Inmueble
Estructura	BATHNUMBER	idealista	Número de baños	Unifamiliar
	BEDROOMNUMBER	idealista	Número de dormitorios	Unifamiliar
	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Unifamiliar
	ENERGYCERTIFICATIONID	idealista	Código de certificado energético	Unifamiliar
	FLATLOCATION	idealista	Indica si el piso es interior o exterior	Plurifamiliar
	FLOOR	idealista	Planta en la que está el inmueble	Plurifamiliar
	GARAGETYPEID	idealista	Tipo de garaje	Unifamiliar
	HASAIRCONDITIONING	idealista	¿Tiene aire acondicionado?	Ambos
	HASANNEX	idealista	El piso tiene anejos	Plurifamiliar
	HASBALCONY	idealista	Tiene balcón	Plurifamiliar
	HASBOXROOM	idealista	Tiene trastero	Unifamiliar
	HASEASTORIENTATION	idealista	Está orientado al este	Unifamiliar
	HASNORTHORIENTATION	idealista	Está orientado al norte	Unifamiliar
	HASPARKINGSPACE	idealista	Tamaño del garaje	Unifamiliar
	HASSOUTHORIENTATION	idealista	Está orientado al sur	Unifamiliar
	HASTERRACE	idealista	Tiene terrazas	Ambos
	HASWARDROBE	idealista	Tiene armarios empotrados	Ambos
	HASWESTORIENTATION	idealista	Está orientado al oeste	Unifamiliar
	ISDUPLEX	idealista	Es un dúplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
	PLOTOFLAND	idealista	Tamaño de la parcela unifamiliar	Unifamiliar
	ROOMNUMBER	idealista	Número de habitaciones	Ambos
	USABLEAREA	idealista	Área útil	Unifamiliar

Fuente: elaboración propia

Debido a que es un modelo de regresión, cuya variable objetivo tiene una gran variabilidad, dispersión y no-normalidad, se opta la modalidad de árboles de

regresión cuantílicos. Adicionalmente, se usan los pesos poblacionales para que el modelo resultante tenga en cuenta la distribución real de la población. Es posible configurar distintos hiperparámetros para mejorar el ajuste de los modelos, tales como:

- *num.trees*: número de árboles utilizados.
- *min.node.size*: tamaño mínimo de nodo final del árbol, en nuestro caso se usa 8 para evitar el sobreajuste.
- *quantreg*: si está activo, el modelo realiza una regresión cuantílica sobre el método *Random Forest* (Meinshausen, 2006), que es una generalización del método original propuesto por Breiman. La ventaja esta clase de regresión es que proporciona información sobre la distribución condicional completa de la variable de respuesta, y no solo sobre la media condicional como en el método de regresión habitual. Su principal ventaha es que este método proporciona una aproximación, no paramétrica y precisa, de regresión cuantílica para problemas con un gran número de variables.
- *importance*: modo de cálculo de la importancia, y representa la forma en la que se mide la capacidad de reducción de la entropía de cada variable. Para este caso, se establece el parámetro *impurity* refierido a la reducción de la varianza.
- *mtry*: número de variables sobre las que se evalúa hacer la división en cada nodo. Por defecto es la raíz cuadrada del número de variables, y nunca debe superar el número total de ellas.
- *maximal.tree.depth*: Indica la profundidad máxima del árbol a construir. En este caso no se restringe este valor.

Dado que existen múltiples combinaciones posibles, se fija *mtry* a la raíz cuadrada del número de parámetros, y se realiza una evaluación de distintas combinaciones del resto de parámetros mediante la técnica de *grid search*[3]. Adicionalmente, se ha aplicado un remuestreo de tipo validación cruzada (LeCun *et al.*, 2015), que ha demostrado ser un criterio consistente para este fin (Yang, 2007).

El criterio de selección de la configuración es aquel que haga máximas las métricas clave (Kuhn *et al.*, 2018), en este caso: el coeficiente de determinación R^2 y el error del modelo, medido como el error cuadrático medio[4] (RMSE). Debido a las diferencias entre los inmuebles de tipo unifamilior y plurifamiliar, se han estimado modelos diferentes por tipología.

Las configuraciones probadas para el modelo ensamblado corresponden al año

[3]La técnica de *Grid Search* consiste en la prueba sistemática de múltiples conmbinaciones de parámetros para un algoritmo de aprendizaje automático. Su objetivo es encontrar la configuración que mejor funciona, en base a unas métricas de rendimiento del modelo.

[4]Tanto el R^2 como el error se miden sobre el conjunto *out of bag*, que ya se analizó en el epígrafe con el mismo nombre del Anexo II, en el capítulo 3.

2019 (se ha tomado este año por ser el más informado de la serie), y se detallan en la Tabla 5.3. Se observa que la precisión aumenta a medida que aumenta aumenta el número árboles y se reduce el tamaño mínimo de nodo. Se decide utilizar 200 árboles con un tamaño mínimo de nodo de 8 instancias. Para unifamiliar, se observa 150 árboles ofrecen un nivel de ajuste similar a 200, sin degradación apreciable en RMSE.

Tabla 5.3. Resultados de la búsqueda de hiperparametros

		Plurifamiliar			Unifamiliar		
num.trees	min.node.size	RMSE	R2	Tiempo	RMSE	R2	Tiempo
200	8	10.34	76.6%	13.84	3.37	81.5%	23.28
150	8	10.44	76.4%	10.39	3.38	81.5%	19.16
100	8	10.60	76.0%	9.68	3.44	81.1%	15.52
50	8	10.95	75.2%	3.69	3.66	79.9%	11.58
200	16	11.55	73.9%	12.59	3.66	79.9%	22.26
150	16	11.63	73.7%	9.88	3.71	79.7%	18.31
100	16	11.77	73.4%	7.83	3.78	79.3%	14.94
50	16	12.05	72.7%	3.58	3.86	78.9%	11.12
10	8	12.94	70.7%	1.60	4.57	75.0%	9.05
200	32	13.19	70.1%	11.85	4.22	76.9%	21.15
150	32	13.25	70.0%	9.46	4.29	76.5%	17.93
100	32	13.41	69.6%	5.91	4.38	76.0%	14.47
50	32	13.66	69.1%	3.48	4.50	75.3%	11.32
10	16	13.85	68.6%	1.54	4.77	73.9%	8.43
10	32	15.18	65.6%	1.51	5.26	71.1%	8.14

Fuente: elaboración propia

Los hiperparámetros seleccionados varían ligeramente según el tipo de inmueble, como muestra la Tabla 5.4. Las diferencias residen en el número de árboles utilizado y el número de variables para las divisiones (*mtry*), en unifamiliar ambos parámetros son ligeramente mayores por su diversidad.

Tabla 5.4. Hiperparámetros modelo de atributos

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	4	TRUE
Unifamiliar	200	impurity	8	5	TRUE

Fuente: elaboración propia

El error se pondera con los elevadores muestrales para priorizar la reducción del error de las observaciones con mayor representación. Por otra parte, aunque no se ha restringido la profundidad máxima del árbol, se ha utilizado un alto número de árboles para limitar el potencial sobreajuste debido al factor anterior.

5.2.2 Modelo hedónico de utilidad de la localización

El uso de información espacial aporta una notable mejora en la calidad del modelado hedónico de índices del precio de la vivienda, y aunque que no hay un método canónico para aplicarla, el aspecto clave es utilizar unidades de análisis suficientemente pequeñas (Hill y Scholz, 2018).

De forma estricta, este modelo no es un modelo hedónico, sino que es un modelo de corrección del sesgo del modelo de atributos, mediante el uso de variables de localización. Por tanto, la variable objetivo del modelo es la proporción entre el precio real por metro cuadrado y la estimación de precio según el modelo de atributos, es decir:

$$y_i = \frac{p_i}{\hat{p_i}^{atributos}} \quad [5.1]$$

donde y_i representa la variable objetivo para la observación i, p_i el precio real y $\hat{p_i}^{atributos}$ el precio estimado por el modelo de atributos.

Como en el resto de casos, se ha utilizado *Random Forests* con los hiperparámetros recogidos en la Tabla 5.5. En este caso, el valor de *mtry* en plurifamiliares es mayor en el modelo de características, por tener un mayor número de covariables.

Tabla 5.5. Hiperparámetros modelo de localización

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	5	TRUE
Unifamiliar	200	impurity	8	5	TRUE

Fuente: elaboración propia

Se han utilizado un conjunto de 18 variables relacionadas con la utilidad marginal asociada a la zona (utilidad/accesibilidad), que se obtienen a través del método presentado en el capítulo 4. Estas variables resumen qué instalaciones o servicios dispone una zona, teniendo en cuenta que los individuos que habitan en cada vivienda se desplazan tanto a pie como en transporte privado.

En la Tabla 5.6, se describen las 32 variables utilizadas, clasificadas en diferentes categorías: accesibilidad, datos básicos del edificio, subtipología y atributos sociodemográficos.

Tabla 5.6. Variables modelo de localización

Categoría	Variable	Fuente	Descripción	Inmueble
Accesibilidad	COMP CAR 1 .. 9	calculado	Accesibilidad en coche	Ambos
	COMP WALK 1 .. 9	calculado	Accesibilidad caminando	Ambos
Area	CLUSTER	calculado	Tipo de zona siguiendo una clasificación propia	Unifamiliar
Edificio	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
Estructura	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Ambos
	ISDUPLEX	idealista	Es un duplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
Mercado	LEADS RESIDENTIAL	idealista	Número de contactos medios en la zona idealista	Plurifamiliar
	ONMARKET RENT	idealista	Número de inmuebles en alquiler en la zona	Plurifamiliar
	ONMARKET SALE	idealista	Número de inmuebles en venta en la zona	Plurifamiliar
	RENTSALE RATIO	idealista	Proporción de número de inmuebles en alquiler versus en compra (en oferta)	Plurifamiliar
Zona	AGE 3	INE	Porcentaje de mayores en la sección censal	Ambos
	DENSITY	INE	Densidad de población del tramo censal	Ambos
	EDUCATION 3	INE	Porcentaje de personas con estudios superiores en la sección censal	Ambos
	POPULATION MUNICIPALITY	INE	Población del municipio	Ambos
	RATE FOREIGN	INE	Tasa de extranjeros en la sección censal	Ambos

Fuente: elaboración propia

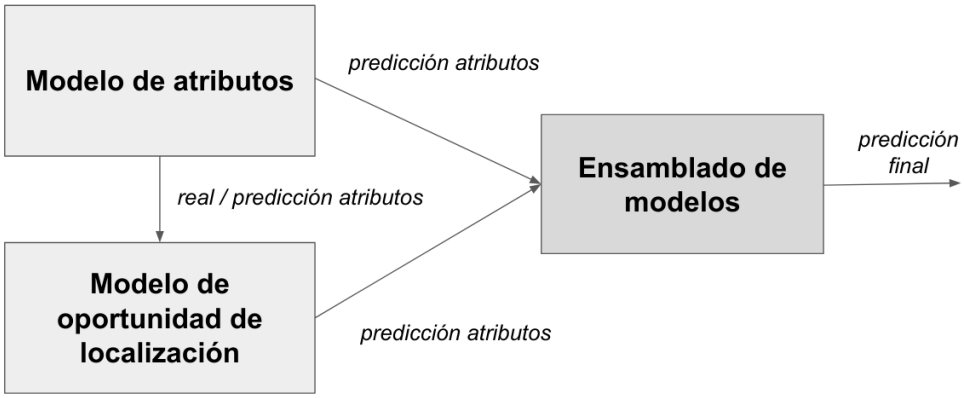
Los atributos zonales se complementan con datos socio-demográficos de la zona y con la tipología de zona calculada (*CLUSTER*). Esto permite explicar los aspectos que no cubren los indicadores de accesibilidad, por ejemplo, dos áreas del norte y sur de Madrid pueden tener precios de suelo diferentes aún teniendo una misma configuración en términos de accesibilidad.

5.2.3 Modelo ensamblado contra modelo único

Se decide construir un modelo final sobre un ensamblado de modelos de características y zona en base a los estudios que demuestran empíricamente que este enfoque produce mejores resultados que los modelos individuales (Graczyk *et al.*, 2009; Hashem, 1997; Krogh y Vedelsby, 1994; Opitz y Shavlik, 1996). En todo caso, es necesario tener en cuenta que, en ocasiones, pueden producirse fenómenos de sesgo (Graczyk *et al.*, 2010).

La forma funcional del ensamblado es semilogarítmica y combina, de forma aditiva, las predicciones del modelos individuales. La idea general, expresada en la Figura 5.2, se basa en que los dos primeros modelos se centran en la reducción de la varianza, cada uno usando un aspecto específico de la información disponible, y el modelo final se enfoca en eliminar el sesgo introducido por cada uno de los anteriores.

Figura 5.2. Esquema general del ensamblado de modelos



Fuente: elaboración propia.

Dado que es posible encontrarse grados específicos de interacción de las variables de cada modelo, porque la influencia del número de habitaciones puede ser distinta entre unos barrios y otros, se usan un conjunto de variables comunes para controlar estas interacciones. De esta forma, se evita que la generación de sesgo, la heterogeneidad espacial o las variables omitidas relevantes.

Los predictores utilizados, se han seleccionado en base a las conclusiones de distintos estudios que los identifican como los más representativos. Véase, por ejemplo, los resultados de Rico y Taltavull (2021) o de Clark et al. (2018). Las

variables utilizadas han sido:

- Tipo de zona: a través de las variables *Cluster* y zona idealista, se utiliza en el modelo final para corregir los sesgos del modelo de localización.
- Superficie útil.
- Estado de la construcción: nueva, segunda mano reformada y segunda mano sin reformar.
- Subtipo de vivienda unifamiliar: solo para esta tipología, necesaria para diferenciar el comportamiento del precio entre adosados, pareados y viviendas independientes.

El uso de atributos zonales en este modelo permite corregir los sesgos espaciales del modelo de localización debidos a variables omitidas. De tal forma que se puede entender que el modelo de localización captura las interacciones generales sobre las áreas cercanas del inmueble, y el modelo ensamblado final, corrige cuestiones desviaciones entre el precio del suelo estimado usando la localización real (se asume como una corrección de sesgo basada en variables ficticias de zona).

Para el último paso, se usan 7 variables, descritas en detalle en la Tabla 5.7, más del subtipo en el caso de vivienda unifamiliar, que permite mejorar la precisión del modelo ensamblado para este caso.

Tabla 5.7. Variables modelo ensamblado

Categoría	Variable	Fuente	Descripción	Inmueble
Modelo	MODEL 1 PREDICTIONS	interno	Predicción modelo de atributos	Ambos
	MODEL 2 PREDICTIONS	interno	Predicción modelo de localización	Ambos
Area	CLUSTER	interno	Tipo de zona - clasificación interna	Ambos
	LOCATIONID	idealista	Código de área idealista	Ambos
Estructura	CONSTRUCTEDAREA	idealista	Área construida	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de unifamiliar	Unifamiliar

Fuente: elaboración propia

A continuación se describe el modelo ensamblado de forma funcional. Para facilitar su lectura se expresa modelo lineal. La siguiente expresión muestra el modelo de atributos, cuya variable dependiente es el logaritmo del precio por metro cuadrado en euros $log(Precio\ m^2)$:

$$log(P_{atributos}) = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot atributo_i + \sum_{k=1}^{n} \beta_k \cdot comunes_i + \varepsilon_a \quad [5.2]$$

donde $atributo_i$ se refiere las covariables de tipo atributo para la observación i, y $comunes_i$ los predictores comunes.

En el modelo de utilidad, las variables independientes principales son los atributos comunes y los atributos del modelo de accesibilidad para cada zona. En este caso, la variable objetivo representa la relación entre el precio real y el precio predicho por el modelo de atributos:

$$R_{utilidad} = \frac{p}{\hat{P}_{atributos}} = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot utilidad_i + \sum_{k=1}^{n} \beta_k \cdot comunes_i + \varepsilon_o \qquad [5.3]$$

donde $utilidad_i$ se refiere las covariables específicas para el modelo de utilidad de la localización.

Finalmente, el modelo ensamblado combina linealmente la contribución de los dos modelos individuales según la expresión:

$$log(P) = \beta_0 + \beta_1 \cdot log(\hat{P}_{atributos}) + \beta_1 \cdot log(\hat{R}_{utilidad}) + \varepsilon_e \qquad [5.4]$$

Los hiperparámetros aplicados al algoritmo *Random Forests*, como muestra la Tabla 5.8, se mantiene casi toda la configuración excepto el parámetro *mtry* debido al reducido número de atributos.

Tabla 5.8. Hiperparámetros modelo ensamblado

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	2	TRUE
Unifamiliar	200	impurity	8	2	TRUE

Fuente: elaboración propia

Para validar la eficacia del modelo ensamblado, se ha desarrollado un modelo que denominaremos único que se tomará como referencia para evaluar el aporte del enfoque ensamblado. El algoritmo utilizado es también *Random Forests*, sobre el que se desarrollan dos modelos diferentes, uno por tipología de vivienda. En ambos casos, la variable a predecir es el logaritmo del precio por metro cuadrado útil, y de forma funcional podría expresarse como:

$$log(P) = \beta_0 + \sum_{i=1}^{n} \beta_i atributos_i + \sum_{i=1}^{n'} \beta'_i \cdot utilidad_i + \sum_{i=1}^{n''} \beta''_i \cdot mercado_i + \varepsilon_e \qquad [5.5]$$

El modelo único cuenta con un mayor número de variables (65), al incorporar información de todos los aspectos del inmueble, como las características

estructurales de la vivienda, los datos de localización o las dinámicas del mercado inmobiliario. El detalle completo de variables utilizadas se presenta en la Tabla 5.9 y la Tabla 5.10.

Tabla 5.9. Variables modelo único

Categoría	Variable	Fuente	Descripción	Inmueble
Accesibilidad	COMP CAR 1 .. 9	calculado	Accesibilidad en coche	Ambos
	COMP WALK 1 .. 9	calculado	Accesibilidad caminando	Ambos
Calidad	CADASTRALQUALITYID	catastro	Calidad de la construcción	Ambos
Edificio	AMENITYID	idealista	Tipo de instalaciones de la finca	Ambos
	BUILTTYPEID	idealista	Nuevo o segunda mano	Ambos
	CHALETTYPEID	idealista	Tipo de inmueble unifamiliar	Unifamiliar
	CONSTRUCTIONYEAR	catastro	Año de construcción de la propiedad	Ambos
	DWELLING COUNT	catastro	Número de inmuebles en la finca	Plurifamiliar
	FLOOR POSITION	idealista	Posición del piso dentro del edificio	Plurifamiliar
	HASDOORMAN	idealista	Tiene portero	Unifamiliar
	HASGARDEN	idealista	Tiene jardín	Unifamiliar
	HASLIFT	idealista	Tiene ascensor	Plurifamiliar
	HASSWIMMINGPOOL	idealista	¿Tiene su edificio una piscina?	Ambos
	MAXBUILDINGFLOOR	idealista	Número de pisos en edificio	Ambos
Fecha	PERIOD	idealista	Fecha en formato YYYYMM	Ambos
Mercado	CHANNELID	idealista	Canal de comercialización	Ambos
	LEADS RESIDENTIAL	idealista	Número de contactos en zona	Ambos
	ONMARKET RENT	idealista	Inmuebles en alquiler en zona	Ambos
	ONMARKET SALE	idealista	Inmuebles en venta	Ambos
	RENTSALE RATIO	idealista	Proporción alquiler/venta	Ambos
Zona	AGE 3	INE	Pct. de mayores	Ambos
	DENSITY	INE	Densidad de población	Ambos
	EDUCATION 3	INE	Pct. personas con estudios superiores	Ambos
	POPULATION MUNICIPALITY	INE	Población del municipio	Ambos
	RATE FOREIGN	INE	Pct. Extranjeros	Ambos

Fuente: elaboración propia

Tabla 5.10. Variables modelo único (continuación)

Categoría	Variable	Fuente	Descripción	Inmueble
Estructura	BATHNUMBER	idealista	Número de baños	Unifamiliar
	BEDROOMNUMBER	idealista	Número de dormitorios	Unifamiliar
	CONSTRUCTEDAREA	idealista	Superficie total en metros cuadrados	Ambos
	ENERGYCERTIFICATIONID	idealista	Código de certificado energético	Unifamiliar
	FLATLOCATION	idealista	Indica si el piso es interior o exterior	Plurifamiliar
	FLOOR	idealista	Planta en la que está el inmueble	Plurifamiliar
	GARAGETYPEID	idealista	Tipo de garaje	Unifamiliar
	HASAIRCONDITIONING	idealista	¿Tiene aire acondicionado?	Ambos
	HASANNEX	idealista	Con anejos (garaje o trastero)	Plurifamiliar
	HASBALCONY	idealista	Tiene balcón	Plurifamiliar
	HASBOXROOM	idealista	Tiene almacenamiento / Trastero	Unifamiliar
	HASEASTORIENTATION	idealista	Está orientado al este	Unifamiliar
	HASNORTHORIENTATION	idealista	Está orientado al norte	Unifamiliar
	HASPARKINGSPACE	idealista	Tamaño del garaje	Unifamiliar
	HASSOUTHORIENTATION	idealista	Está orientado al sur	Unifamiliar
	HASTERRACE	idealista	Tiene terrazas	Ambos
	HASWARDROBE	idealista	Tiene armarios empotrados	Ambos
	HASWESTORIENTATION	idealista	Está orientado al oeste	Unifamiliar
	ISDUPLEX	idealista	Es un duplex	Plurifamiliar
	ISPENTHOUSE	idealista	Es un ático	Plurifamiliar
	ISSTUDIO	idealista	Es un estudio	Plurifamiliar
	PLOTOFLAND	idealista	Tamaño de la parcela unifamiliar	Unifamiliar
	ROOMNUMBER	idealista	Número de habitaciones	Ambos
	USABLEAREA	idealista	Área útil	Unifamiliar

Fuente: elaboración propia

Los hiperparámetros aplicados al algoritmo *Random Forests* son similares a los usados en los ensamblados, como muestra la Tabla 5.11. Con la excepción del parámetro *mtry* que se establece en 7, debido al mayor número de variables utilizadas.

Tabla 5.11. Hiperparámetros modelo ensamblado

Inmueble	Árboles	Importancia	Tamaño nodo	Mtry	Quantreg
Plurifamiliar	150	impurity	8	7	TRUE
Unifamiliar	150	impurity	8	7	TRUE

Fuente: elaboración propia

5.3 Resultados

La metodología hedónica de oferta tiene como objetivo producir un modelo general y preciso, por tanto, se evaluará la capacidad de generalización y ajuste de los modelos construidos. Adicionalmente, se estudiará si la incorporación del enfoque ensamblado produce los beneficios esperados, junto con el papel que juegan las distintas variables en la explicación del precio.

Sobre los 36 modelos finales construidos[5], se estudia cuál es la aproximación que mejor resultados ofrece utilizar (ensamblado o único), a través de su grado de ajuste, forma de los errores de predicción y distribución geográfica de los mismos.

5.3.1 Medidas para evaluar la calidad de los modelos

Existe un amplio número de métricas orientadas disponibles para evaluar la calidad de un modelo de precios de la vivienda. Steurer *et al.* (2021) identifican 49 medidas, de las que recomiendan 7, que cubren varios aspectos de análisis sobre los residuos: sesgo, desviación absoluta, ratio de desviación absoluta, desviación cuadrática, ratio sobre desviación cuadrática y porcentaje de errores fuera de rango, medidas sobre cuantiles. Sobre esta base, se han tomado las métricas recogidas en la Tabla 5.12, complementadas con otras de ajuste espacial y sesgo (presentadas en el epígrafe 3.4). El número de medidas debe ser abundante para asegurar la robustez de los resultados (Steurer *et al.*, 2021).

Tabla 5.12. Medidas de error de modelos

Medida	Fórmula
Raíz cuadrada del error medio cuadrático medio	$RMSE = \frac{1}{N}\sum_{t=1}^{N}\sqrt{e_i^2}$
Raíz cuadrada del error medio cuadrático medio normalizado	$NRMSE = \frac{1}{\overline{x}}\frac{1}{N}\sum_{i=1}^{n}\sqrt{e_i^2}$
Error medio absoluto	$MAE = \frac{1}{N}\sum_{i=1}^{N}e_i^2$
Error medio absoluto	$MAE = \frac{1}{N}\sum_{i=1}^{N}e_i^2$
Error mediano absoluto	$MedAE = median\|e_i\|$
R^2	$R^2 = 1 - \frac{\sigma_e^2}{\sigma^2}$
Error medio en porcentaje	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\|e_i\|}{x_i}$
Error mediano en porcentaje	$MEDAPE = mediana\frac{\|e_i\|}{x_i}$

Fuente: elaboración propia

e_i representan los errores de estimación y N el tamaño muestral

[5] Los 36 modelos proceden de las combinaciones entre modelo ensamblado y único, los dos tipos de vivienda y 9 años.

El estudio del ajuste de los modelos se ha realizado en términos monetarios, en linea con la investigación de Pérez-Rave *et al.* (2019), que indica que el análisis de errores sobre la variable transformada[6], a menudo, puede ofrecer resultados ficticios desde un punto de vista estadístico.

Adicionalmente, el sesgo de los modelos estimados representa de la desviación entre la esperanza de los valores predichos y los observados[7]. De forma intuitiva, se puede calcular como la media de las desviaciones de las estimaciones del modelo, que en este caso se realiza de varias formas: como la desviación en términos absolutos, en porcentaje y como curtosis de la distribución de precios estimados. En particular, se ha utilizado la siguiente expresión:

$$Bias(\hat{f}(x)) = E[\hat{f}(x)] = \hat{f}_{observado}(x) - \hat{f}_{estimado}(x) \quad [5.6]$$

donde $\hat{f}(x)$ un estimador sobre el conjunto s, y E representa la esperanza de la desviación de dicho estimador, entre datos observados y predichos.

Todas las métricas de ajuste y error utilizadas, se han calculado sobre las muestras denominadas OOB de los modelos *Random Forests*. Para más información véase el Anexo 3b del capítulo 3.

5.3.2 Resultados de ajuste del modelo

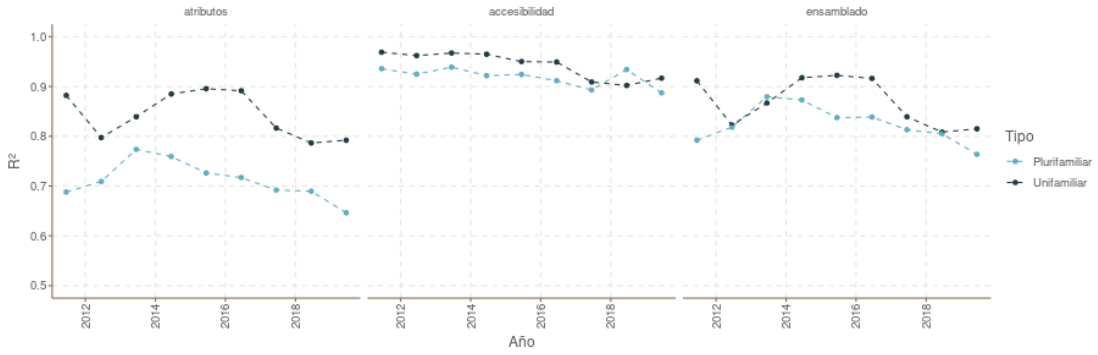
El planteamiento ensamblado pretende que los modelos individuales (atributos y localización) reduzcan la varianza lo máximo posible y, que al combinarlo, se ajusten sus sesgos particulares.

La Figura 5.3 muestra como se comporta el coeficiente de determinación R^2 de los tres procesos, observándose que los modelos individuales capturan buena parte del espectro de la varianza. Con el objetivo de facilitar la correcta interpretación de las medidas de ajuste de los modelos, se recuerda que el coeficiente de determinación del modelo de utilidad se refiere al ajuste sobre razón entre el precio real y el predicho por el modelo de atributos.

[6] A través de una transformación logarítmica o potencial, como puede ser el logaritmo del precio por superficie útil.

[7] El estimador al que se refiere puede ser el precio de la vivienda, los errores del modelo u otra medida de interés relacionada con el modelo.

Figura 5.3. Niveles de ajuste por modelos



Fuente: elaboración propia.

El ajuste medio del modelo de atributos es más bajo debido a que la ubicación geográfica exacta es un determinante importante del precio, en todo caso, el ajuste es satisfactorio al incorporar una mínima especificación zonal, procedente del atributo código de clúster.

La Tabla 5.13 indica que el comportamiento del R^2 es relativamente estable en el tiempo, aunque con cierta degradación en 2012. Se observa que, a partir de 2016, el peor rendimiento del modelo de atributos se corrige eficazmente con el ensamblado. Además, en general, los niveles de ajuste son más altos en unifamiliar que en plurifamiliar, probablemente a consecuencia de una mayor heterogeneidad de este tipo de vivienda.

Tabla 5.13. Ajuste de los modelos en R² desglosado por tipologías

Tipo	Modelo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Unifamiliar	atributos	88.2%	79.7%	83.9%	88.5%	89.6%	89.1%	81.6%	78.7%	79.2%
	accesibilidad	96.9%	96.2%	96.7%	96.5%	95.0%	94.9%	90.9%	90.2%	91.7%
	ensamblado	91.1%	82.2%	86.7%	91.8%	92.2%	91.6%	83.9%	80.8%	81.5%
Plurifamiliar	atributos	68.8%	70.9%	77.4%	75.9%	72.6%	71.7%	69.2%	69.0%	64.6%
	accesibilidad	93.6%	92.5%	93.9%	92.2%	92.4%	91.2%	89.3%	93.4%	88.7%
	ensamblado	79.2%	81.8%	87.9%	87.3%	83.7%	83.9%	81.3%	80.5%	76.4%

Fuente: elaboración propia

Desde un punto de vista comparativo, sobre la medida de R^2, el modelo muestra un rendimiento en un orden similar con otros modelos de precio de alquiler de la literatura, con valores (caso ensamblado) entre 0,76 y 0,92. Aunque es necesario recordar que la escala de esta medida depende mucho de varios factores como la fuente de la calidad de la información o el algoritmo utilizado, se toman como referencia los casos de: Chung (2015) que presenta un 0,98, con un modelo

semilogarítmico; Löch (2010) consigue un 0.856; Fuss y Koller (2016) un 0,883, y Clark y Lomax (2018) logran un 0,69.

En términos de error, la Tabla 5.14 muestra errores medios crecientes en los dos tipos de vivienda, en parte debido a que los precios son crecientes. En el valor normalizado, las viviendas plurifamiliares han un error estable en torno al 17%, mientras que, las unifamiliares ofrecen errores con mucha más fluctuación.

Tabla 5.14. RMSE (absoluto) y NRMSE (normalizado), modelo ensamblado

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Absoluta	Plurifamiliar	28,13	24,59	18,84	20,04	25,30	29,18	34,43	34,54	38,77
	Unifamiliar	15,00	20,87	16,61	12,19	11,61	12,36	17,78	21,70	22,04
Normalizada	Plurifamiliar	17.0%	15.7%	12.8%	13.5%	16.3%	17.3%	17.7%	16.0%	17.3%
	Unifamiliar	14.3%	20.8%	17.9%	13.4%	12.6%	12.6%	17.3%	19.7%	19.2%

Fuente: elaboración propia

Aunque no son medidas totalmente comparables, el *NRMSE* y el *MAPE* guardan relación al ser medidas de error tipificadas, siendo la primera mucho más exigente ante errores extremos, y por tanto ofrece datos más altos, algo que se puede confirmar más adelante en la Tabla 5.19.

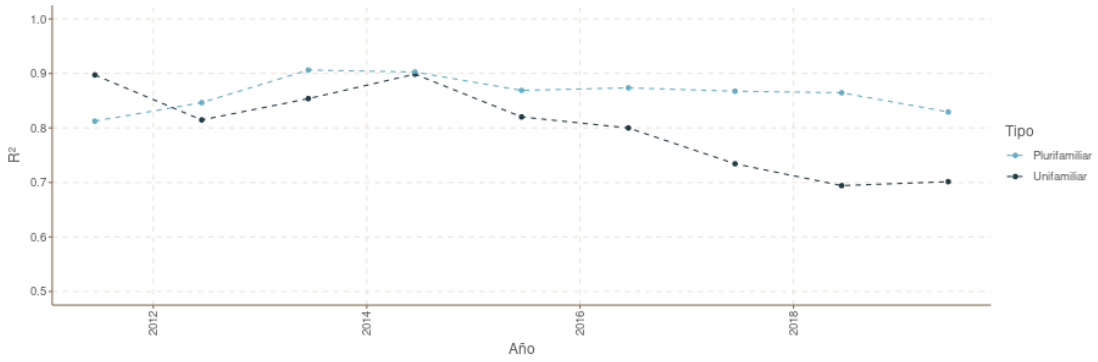
Cuando se compara el *NRMSE* de nuestro modelo con el *MAPE* del estudio de Alfaro *et al.* (2020), se observa como en nuestro caso los valores se encuentran en un orden de magnitud similar al *MAPE* medio municipal[8] de *Random Forests* para este estudio (15,93). Asimismo, este valor tampoco es muy diferente a los valores del modelo desarrollado por Clary y Lomax (2018), en cuyo caso, el *MEDAPE* es aún más optimista que el *MAPE*.

5.3.2.1 Selección de modelo: ensamblado o único

El modelo único muestra un comportamiento de ajuste decreciente, como se aprecia en la Figura 5.4, con una degradación más acusada en las viviendas unifamiliares a partir de 2015.

[8] Sobre 63 municipios españoles de más de 100.000 habitantes.

Figura 5.4. Niveles de ajuste del modelo único, desglosado por tipo



Fuente: elaboración propia.

El ajuste en las viviendas plurifamiliares se mantiene estable en el tiempor, como se aprecia en la Tabla 5.15.

Tabla 5.15. Ajuste de los modelos desglosados por tipologías (modelo único)

Tipo	Modelo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Unifamiliar	unico	89.7%	81.5%	85.4%	89.9%	82.0%	80.0%	73.4%	69.4%	70.1%
Plurifamiliar	unico	81.2%	84.6%	90.6%	90.3%	86.9%	87.4%	86.7%	86.5%	82.9%

Fuente: elaboración propia

El modelo único muestra errores cuadráticos menores en las viviendas plurifamiliares, con un mayor grado de estabilidad temporal en los errores, tal y como se ve en la Tabla 5.16. En cambio, para las viviendas unifamiliares, esta versión ofrece peores errores que el modelo agregado.

Tabla 5.16. RMSE (absoluto) y NRMSE (normalizado), modelo único

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
Absoluta	Plurifamiliar	26,71	22,60	16,61	17,54	22,73	25,82	29,00	28,80	32,94
	Unifamiliar	16,16	21,33	17,40	13,52	18,85	20,63	24,36	29,55	30,26
Normalizada	Plurifamiliar	16.1%	14.4%	11.2%	11.8%	14.7%	15.3%	14.9%	13.4%	14.7%
	Unifamiliar	15.4%	21.2%	18.7%	14.9%	20.4%	21.1%	23.7%	26.8%	26.4%

Fuente: elaboración propia

La Tabla 5.17 compara la precisión del modelo ensamblado con respecto el único, a través de los errores y el R^2 en la muestra OOB. Se aprecia una notable diferencia en los resultados en función de la tipología. Por una parte, en las viviendas plurifamiliares el modelo único es mejor que el ensamblado en todos los periodos; y por otra, del caso de las unifamiliares, donde se produce lo contrario.

En términos de magnitud, también se aprecian diferencias, siendo las mejoras el doble de grandes en las plurifamiliares que en las unifamiliares.

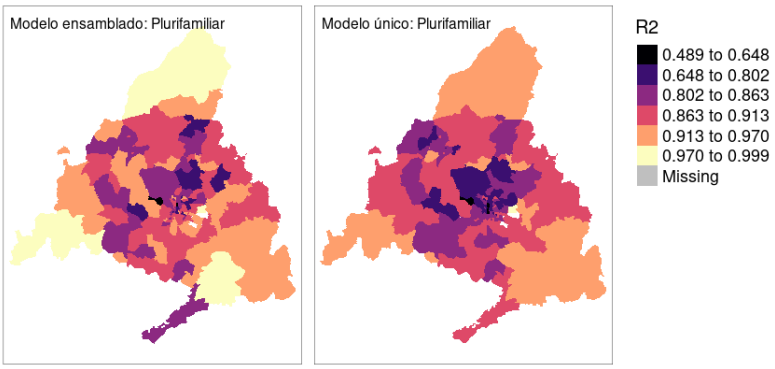
Tabla 5.17. Mejora (+) o empeoramiento (-) del modelo ensamblado, RMSE y R^2

Métrica	Tipo	2011	2012	2013	2014	2015	2016	2017	2018	2019
RMSE	Plurifamiliar	-1,42	-1,99	-2,23	-2,51	-2,58	-3,36	-5,42	-5,74	-5,82
	Unifamiliar	1,16	0,45	0,80	1,33	7,24	8,27	6,58	7,85	8,22
R^2	Plurifamiliar	-2.0%	-2.8%	-2.7%	-3.0%	-3.1%	-3.5%	-5.4%	-5.9%	-6.6%
	Unifamiliar	1.4%	0.8%	1.3%	1.9%	10.2%	11.6%	10.5%	11.4%	11.4%

Fuente: elaboración propia

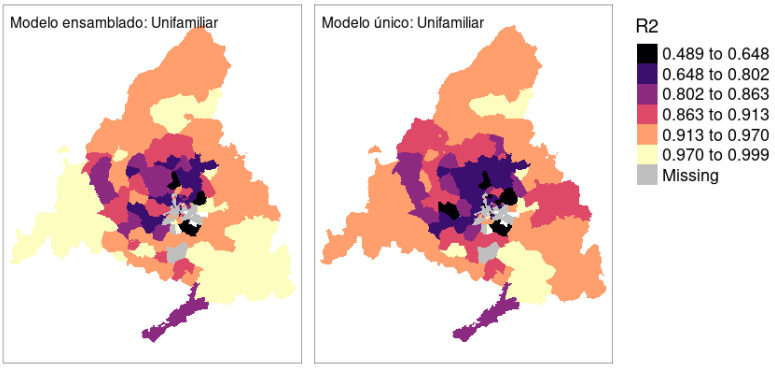
Dado que es deseable tener tanto un buen ajuste global como para las distintas zonas geográficas de análisis, se muestra el aspecto de ajuste zonal[9] en la Figura 5.6 y la Figura 5.5. Se aprecia que los modelos ensamblados mejoran al único resultados para todas las zonas, en los dos tipos de vivienda.

Figura 5.5. Ajuste espacial en vivienda plurifamiliar, Comunidad de Madrid



Fuente: elaboración propia.

Figura 5.6. Ajuste espacial en vivienda unifamiliar, Comunidad de Madrid

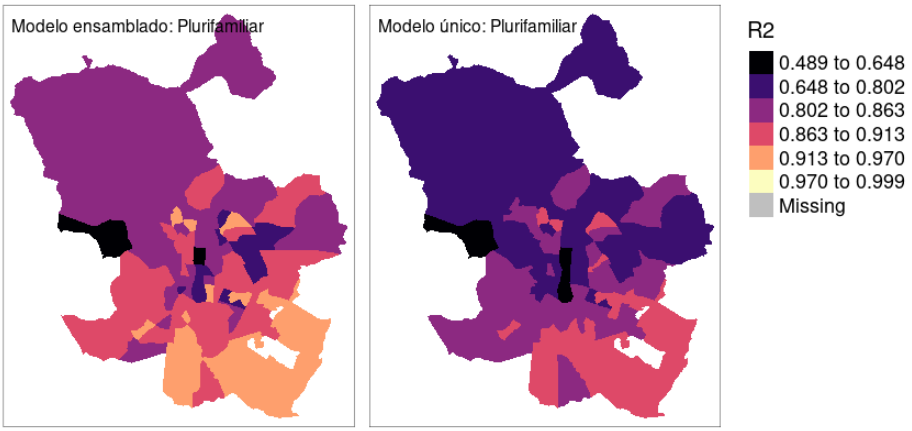


Fuente: elaboración propia.

[9] Datos calculados sobre todos los periodos desde 2011 a 2019, mediante validación cruzada (K=10).

Para el municipio de Madrid, se analiza sólo el ajuste para plurifamiliares, dada la poca representatividad de las unifamiliares. La Figura 5.7 se aprecia como los niveles de ajuste son menores en el modelo único, y con un alto grado de heterogeneidad zonal, siendo las regiones con mayor precisión las ubicadas en el sur y este de la ciudad. Por contra, las zonas con peor ajuste, son las el centro y el eje de la Castellana, debido a que en estas áreas existe una mayor heterogeneidad en precios y características.

Figura 5.7. Ajuste espacial en vivienda plurifamiliar, municipio de Madrid



Fuente: elaboración propia.

Desde un punto de vista numérico, los datos de las figuras anteriores se resumen en la Tabla 5.18, desglosados por tipo aproximación, tipo de vivienda y clase de zona. Se aprecia que los errores del modelo y coeficientes de determinación disminuyen siempre en los modelos ensamblados.

Tabla 5.18. Ajuste espacial promedio por clase de modelo y tipo

		Ensamblado		Único	
Capital	Tipo	R^2	RMSE	R^2	RMSE
No	Plurifamiliar	86,0%	1.29	79,9%	1.57
	Unifamiliar	80,6%	0.89	71,5%	1.22
Sí	Plurifamiliar	89,3%	0.67	87,1%	0.76
	Unifamiliar	91,3%	0.48	88,8%	0.59
Todos	Plurifamiliar	87,2%	1.06	82,5%	1.27
	Unifamiliar	85,9%	0.71	80,1%	0.94

Fuente: elaboración propia

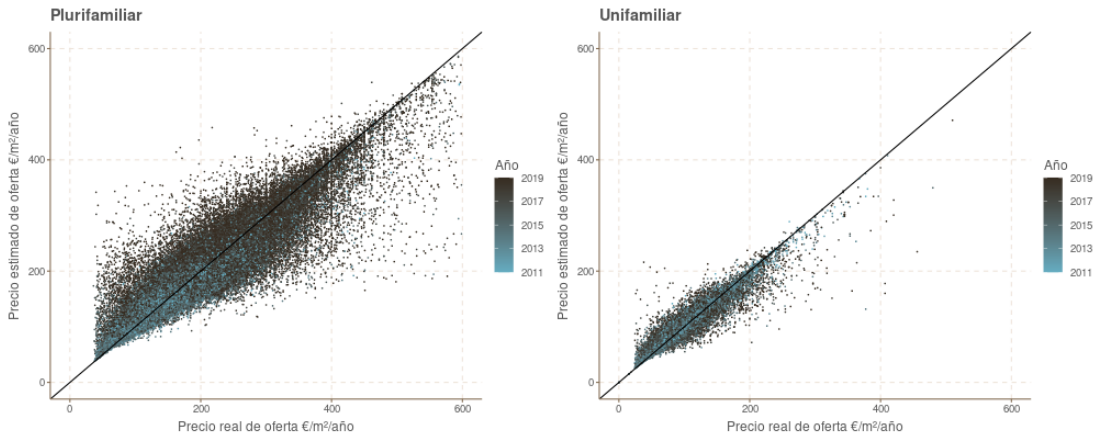
Los resultados globales y desglosados por zona contrapuestos se podrían

considerar un caso particular de la paradoja de Simpson[10] (Wagner, 1982). Esta discrepancia se debe a la ausencia de control de una variable de confusión relevante, que en este caso podría ser la geografía, cuya falta introduciría sesgos por la heterogeneidad espacial de los precios (McLaughlin y Young, 2018), o por un desigual nivel de información de los estratos que componen la muestra (Boeing, 2020). Ante esta divergencia de resultados, y debido a que el análisis geográfico al ser más completo potencialmente ofrece resultados más robustos (Steurer *et al.*, 2021), se opta por tomar el modelo que mejor funciona en este caso, es decir, el ensamblado.

5.3.2.2 Métricas sobre la población total (*in-bag*)

El desglose de las medidas, obtenidas sobre la población OOB es reducido, por una limitación técnica del algoritmo utilizado. Para disponer de un análisis en detalle, se han calculado una serie de métricas sobre la población completa (*in-bag*), que no pretenden estudiar la capacidad de generalización del modelo, sino las métricas clave, errores y sesgos de los diferentes estratos de la población. Un segundo motivo es que los precios estimados sobre el conjunto *in-bag*, se usan como base del índice de precios de oferta y como entrada del modelo de conversión definitivo. Por tanto, es necesario estudiar las diferencias entre los valores reales y los inferidos por el método. Para reducir sesgos de selección, en todas las medidas, se utiliza un remuestreo del tipo validación cruzada con 10 mezclas (K=10).

Figura 5.8. Ajuste de la regresión, precio de oferta



Fuente: elaboración propia.

La calidad del ajuste del modelo se representa en la Figura 5.8. La línea de 45º representa el ajuste perfecto, por tanto, cualquier desviación sobre ella indica un error de estimación del precio. Se aprecia que el segmento de viviendas

[10] La paradoja de Simpson se produce cuando el agregado de una medida sobre varias categorías muestra una incidencia mayor o menor al de cualquiera de las categorías individuales.

plurifamiliares tiene mayores desviaciones, especialmente los últimos años de la serie. Del mismo modo, los rangos de precios más altos son menos precisos, probablemente porque implican un mayor grado de heterogeneidad no controlada en el hedónico, bien debida a variables ausentes, o bien por falta de soporte de datos por ser un segmento muy minoritario.

La Tabla 5.19 muestra las medidas de error para viviendas plurifamiliares ponderadas por los factores de elevación. Se observa que el error cuadrático (RMSE) mantiene en órdenes de magnitud similares a los obtenidos en la métricas OOB, lo que indica que el modelo generaliza correctamente.

Tabla 5.19. Métricas in-bag de los modelos, viviendas plurifamiliares

Año	N	RMSE	NRMSE	MAE	MAPE	MEDAE	MEDAPE
2011	206.969	13.76	9,49%	7.72	5,32%	3.99	2,75%
2012	361.230	13.22	9,55%	7.40	5,35%	3.96	2,86%
2013	491.120	12.14	9,21%	6.63	5,03%	3.37	2,56%
2014	495.818	12.74	9,80%	6.64	5,11%	3.12	2,40%
2015	431.065	14.18	10,38%	7.10	5,20%	3.09	2,26%
2016	355.787	16.14	11,25%	8.06	5,62%	3.26	2,27%
2017	461.683	16.10	10,06%	8.00	5,00%	3.10	1,94%
2018	620.972	14.35	8,33%	7.01	4,07%	2.63	1,53%
2019	705.461	15.98	9,05%	7.34	4,16%	2.52	1,43%

Fuente: elaboración propia

Los errores porcentuales, tanto medios como medianos, son satisfactorios a la luz de lo comentado en el epígrafe 5.3.2. Además, se confirma que los errores cuadráticos normalizados son menores a los porcentuales absolutos, lo mismo que los valores medianos son siempre menores a los medios.

Los errores del segmento unifamiliar, recogidos en la Tabla 5.20, muestran valores muy bajos, aunque mayores que los presentados anteriormente para el conjunto OOB. Esto se puede atribuir a dos motivos: 1) las medidas OOB no se calculan ponderadas, y 2) que un menor tamaño poblacional y geográficamente más amplio, puede impactar en fenómeno de heterogeneidad espacial de los precios, que se sugería en el epígrafe 5.3.2.1, y mostrado gráficamente en la Figura 2.12 del capítulo 2.

Tabla 5.20. Métricas in-bag de los modelos, viviendas unifamiliares

Año	N	RMSE	NRMSE	MAE	MAPE	MEDAE	MEDAPE
2011	17.159	4.83	5,27%	1.99	2,18%	0.50	0,55%
2012	32.863	4.58	5,15%	1.82	2,05%	0.47	0,52%
2013	52.297	3.91	4,83%	1.41	1,74%	0.31	0,39%
2014	54.441	3.81	4,97%	1.30	1,70%	0.30	0,39%
2015	40.395	3.49	4,35%	1.15	1,44%	0.26	0,32%
2016	33.628	3.50	4,21%	1.16	1,40%	0.27	0,33%
2017	31.135	3.88	4,54%	1.13	1,32%	0.26	0,30%
2018	31.525	4.05	4,41%	1.18	1,29%	0.30	0,33%
2019	30.950	4.15	4,41%	1.19	1,26%	0.32	0,34%

Fuente: elaboración propia

Las métricas de error, desglosadas en función de si la vivienda se encuentra en la capital o el resto de la provincia, se recogen en la Tabla 5.21. Se observa que los errores en Madrid son más altos que los del resto de la Comunidad. Esta diferencia se debe a factores: el primero, que la proporción de viviendas unifamiliares es mayor en las zonas rurales y el extrarradio, como vemos en la Figura 2.12 del epígrafe 2.4.3; y el segundo, que la diversidad de zonas es mucho mayor en la ciudad, lo que implica una mayor dificultad en la especificación de los modelos, y por tanto, un mayor grado de errores de estimación.

Tabla 5.21. Métricas in-bag, capital o resto de provincia

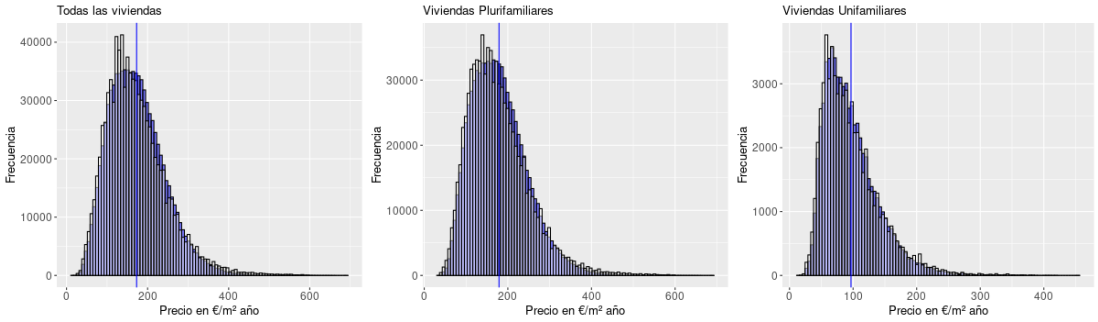
	Madrid						Resto					
Año	N	RMSE	MAE	MAPE	MEDAE	MEDAPE	N	RMSE	MAE	MAPE	MEDAE	MEDAPE
2011	159.508	16.27	9.97	6,02%	5.95	3,59%	64.620	7.62	3.73	3,44%	1.83	1,69%
2012	270.471	15.85	9.67	6,04%	5.88	3,67%	123.622	7.30	3.70	3,58%	1.91	1,85%
2013	358.096	15.29	9.33	5,99%	5.69	3,66%	185.321	5.87	3.01	3,04%	1.49	1,50%
2014	365.806	16.27	9.77	6,25%	5.81	3,72%	184.453	6.27	2.78	2,88%	1.22	1,27%
2015	319.303	18.81	11.02	6,62%	6.36	3,82%	152.157	6.04	2.73	2,68%	1.14	1,12%
2016	266.035	21.52	12.75	7,22%	7.25	4,10%	123.380	6.65	2.89	2,74%	1.15	1,09%
2017	374.100	21.33	12.64	6,48%	6.87	3,52%	118.718	6.22	2.65	2,26%	1.00	0,85%
2018	530.889	19.37	11.20	5,37%	5.85	2,80%	121.608	5.25	2.47	1,91%	0.97	0,75%
2019	601.289	20.45	11.14	5,24%	5.35	2,52%	135.122	9.35	3.37	2,46%	1.11	0,81%

Fuente: elaboración propia

5.3.3 Análisis de sesgos y residuos

Como se mencionaba en la metodología, los modelos ensamblados reducen de forma eficaz de reducir la varianza de los errores, pero pueden mostrar sesgos (Graczyk *et al.*, 2010). Se estudia el sesgo de los modelos a través de sus residuos, con una estrategia de remuestreo de tipo validación cruzada ponderada utilizando los pesos poblacionales. La Figura 5.9 muestra el sesgo de estimación a través de la función de densidad ponderada[11]. Se observa la existencia existe de un ligero sesgo positivo del error, que es mayor para las viviendas plurifamiliares.

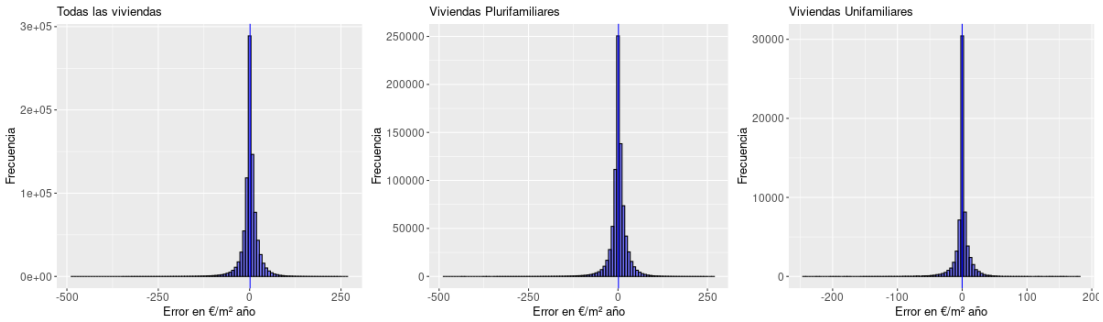
Figura 5.9. Histograma de precios reales (blanco) versus precios estimados por el modelo (azul), totales y desglosados por tipo de vivienda



Fuente: elaboración propia.

Los residuos muestran una alta concentración en torno al cero, como se observa en la Figura 5.10. Dichos errores se distribuyen de forma simétrica con respecto al origen, con un rango más amplio en los errores negativos que en los positivos. Por consiguiente, se deduce que los modelos tienden a infraestimar los casos más extremos, aunque ofrecen un buen comportamiento en el grueso de la población.

Figura 5.10. Distribución de estimación, totales y desglosados por tipo de vivienda



Fuente: elaboración propia.

Para analizar el grado de variabilidad de los errores, se ha calculado el coeficiente

[11] Se han utilizado los pesos poblacionales para ajustar la frecuencia de los valores en lugar de construir la función de densidad sobre las observaciones.

de dispersión ponderado (*COD*) sobre el cociente de la estimación sobre el valor real. Se toma la definición de la medida propuesta por Steurer *et al.* (2021) adaptada a una población ponderada, calculada según la siguiente expresión analítica:

$$COD = \frac{1}{W}\sum_{i=1}^{N}\left| w_i \frac{\hat{p_i}}{p_i} / \left(\sum_{i=1}^{N} w_i \left(\frac{\hat{p}}{p}\right)\right) - 1\right| \quad [5.7]$$

donde W es la suma de los pesos muestrales w_i, para cada observación i, con un precio real $\hat{p_i}$ y uno estimado $\hat{p_i}$.

En términos numéricos, resumidos en la Tabla 5.22, se observa que los errores de las viviendas de tipo plurifamiliar son positivos en todos los periodos, con un incremento gradual en el tiempo. Este comportamiento se puede generalizar para todo el espectro de valores, a la vista de los valores de los cortes cuantílicos y el rango intercuartílico (IQR)[12]. El COD, inferior al 5%, se encuentra en rangos muy inferiores a otras referencias como el de Steurer *et al.* (2021) con un 23,3%, o el de Alfaro (2020), con un valor medio de 15,56%, si bien, en ambos casos los modelos se referían precios de venta en vez de alquiler.

Tabla 5.22. Distribución error en €/m²/año, vivienda unifamiliar

Año	Media	P05	P25	P50	P75	P95	IQR	COD
2011	-0.44	-20.63	-4.05	0.03	4.18	18.34	8.23	2.9%
2012	0.05	-25.86	-6.02	0.42	7.32	25.23	13.34	5.0%
2013	-0.53	-19.03	-3.79	0.01	3.88	16.24	7.66	2.9%
2014	-0.69	-22.72	-4.09	0.04	4.22	18.68	8.32	2.8%
2015	-0.91	-27.27	-4.78	0.02	4.89	21.98	9.67	3.0%
2016	-1.14	-32.75	-5.96	0.03	5.95	26.65	11.91	3.2%
2017	-1.09	-40.42	-6.84	0.05	7.66	33.79	14.50	3.0%
2018	-0.84	-42.64	-6.51	0.04	7.51	37.08	14.02	2.4%
2019	-0.74	-49.22	-7.72	0.04	9.01	43.95	16.73	2.3%

Fuente: elaboración propia

En unifamiliares, los errores son mucho menores (aproximadamente la mitad que en plurifamiliar), y toman tanto valores negativos y positivos, como muestra la Tabla 5.23. La una distribución es relativamente simétrica (en función de la diferencia entre el P05 y el P95), con una diferencias de signo entre la mediana y la media, que indica que el modelo es más impreciso en las predicciones de los precios altos más extremos. El coeficiente de dispersión, también se encuentra en valores muy bajos y estables en el tiempo.

[12] El rango intercuartílico se calcula como la diferencia en valor absoluto del valor del percentil 75 y el percentil 25.

Tabla 5.23. Distribución error en €/m²/año, vivienda unifamiliar

Año	Media	P05	P25	P50	P75	P95	IQR	COD
2011	0.20	-9.94	-1.72	0.07	2.06	11.04	3.78	2.3%
2012	0.08	-9.25	-1.59	0.03	2.09	10.05	3.68	1.9%
2013	0.04	-10.59	-1.66	0.03	1.97	10.74	3.63	1.8%
2014	0.15	-12.80	-1.63	0.04	2.30	13.77	3.93	1.8%
2015	0.21	-14.07	-1.77	0.00	2.42	15.91	4.19	1.4%
2016	0.21	-15.96	-1.87	0.04	2.83	15.66	4.70	1.4%
2017	-0.43	-28.18	-3.59	0.06	5.55	22.88	9.14	1.4%
2018	-0.13	-33.98	-4.12	0.24	6.78	29.02	10.90	1.4%
2019	-1.13	-33.77	-5.17	0.08	6.05	26.27	11.22	1.3%

Fuente: elaboración propia

5.3.4 Interpretabilidad, importancia de variables

Hasta ahora se ha analizado el funcionamiento del modelo en términos de ajuste, pero no se ha entrado en la contribución de las distintas variables, que es la base en la que se sustenta la teoría de precios hedónicos. Aunque los modelos de árboles han sido denominados históricamente como “cajas negras”, porque no generan coeficientes para cada característica (Rico y Taltavull, 2021), existen mecanismos que permiten interpretar las contribución en términos de importancia[13]. De forma general, hay 3 formas de evaluar la importancia de las características, con respecto a los poderes predictivos del modelo:

- De filtro: las características se filtran independientemente del modelo a través de criterios en sus propias propiedades (correlación con el objetivo
- De envoltorio (*wrapper*): basados en algoritmos de búsqueda que tratan a los predictores como entradas y utilizan el error como la salida a optimizar.
- Integradas: la selección de características integradas combinan algoritmos de búsqueda del predictor con la estimación de parámetros y, generalmente, se optimizan con una única función objetivo.

La estimación de la importancia de las variables en un modelo de ensamblado se debe realizar de forma diferente a la de un modelo de regresión por mínimos cuadrados, como describe Grömping (2009). Para el caso de *Random Forests*, los dos métodos más comunes son el de impureza y el de permutación. El de impureza mide la contribución de la variable en la reducción de la medida de desorden del modelo. El de permutación mide el impacto que tendría en la precisión si se permutan los valores entre todos los registros de la muestra[14].

[13] La importancia de cada variable es una medida que relaciona la capacidad marginal de cada independiente de explicar la variable dependiente.

[14] Se mide el efecto de usar el valor real, sobre un valor aleatorio que sigue la misma distribución

La forma más común de determinar la importancia de las variables es la impureza (Strobl *et al.*, 2007), basada en el criterio usado internamente para la determinación de los cortes del árbol de decisión. Cada uno de los cortes son una condición que se aplica a una sola característica (ejemplo: área mayor de 100 metros), y subdivide el conjunto de datos en dos grupos, de forma que la entropía de la variable respuesta disminuye en ambos subconjuntos, o dicho de otro modo, divide la población en los dos grupos lo más homogéneos posible. Para estimar el criterio de corte, se elige la condición sobre una variable localmente[15] la reducción de la varianza, la cual es mayor para los precios de cualquiera de los grupos resultantes que para el conjunto original.

Puesto que se conocen el impacto que tiene el corte sobre una variable, en la reducción de impureza para cada uno de los cortes, la disminución de impurezas de cada característica se calcularía como el un promedio ponderado de estas medidas.

La importancia se ha calculado de forma general para todo el modelo, por tanto, no es posible determinar las contribuciones marginales para cada registro predicho. No obstante, existen métodos más sofisticados, como los basados en Shapley (Roth, 1988), que aproximan la importancia de una solución según el reparto de coste y beneficios entre los colaboradores, en este caso los predictores. Dicho de otra manera, la esperanza de la contribución marginal de cada predictor, para todos los casos posibles (Winter, 2002). Este método ha evolucionado en los últimos años en un marco que une la idea original y la teoría de juegos, denominado SHAP[16] Lundberg y Lee (2017) y Lundberg *et al.* (2018) desarrollan en profundidad los últimos avances en este método aplicado a ensamblados de árboles. A modo de ejemplo, Rico y Taltavull (2021) desarrollan un análisis en profundidad de los valores SHAP aplicados a un modelo de precios de la vivienda en Alicante.

En este caso, la medida de importancia se ha tipificado a valores entre 0 a 100, en la que 100 representa el mayor grado de importancia por variable. Puesto que el cálculo de la importancia de las variables para el modelo ensamblado es complejo de intepretar, se estudiarán las contribuciones de las variables para cada una de las partes del ensamblado.

Para el modelo de vivienda plurifamiliar, los atributos de superficie, número de habitaciones tamaño y calidad del edificio son los más relevantes, como se ver en la Tabla 5.24. A pesar de que las medidas son difícilmente comparables

[15] Se indica que es localmente porque la evaluación del impacto sobre la entropía se realiza para ese corte no para la población general del árbol.

[16] SHapley Additive exPlanations (SHAP) es un método que permite calcular, a nivel de observación individual, el grado de contribución de las distintas variables de entrada en el resultado inferido por un modelo de aprendizaje automático.

de estudio en estudio, este resultado es consistente con la presencia en las primeras posiciones en importancia de las variables de otros trabajos, dando mayor prioridad a aquellas que contienen la superficie, antigüedad o número de habitaciones (Clark y Lomax, 2018; Füss y Koller, 2016; Hanink *et al.*, 2012; Rico y Taltavull, 2021).

Tabla 5.24. Importancia de variables plurifamiliar modelo de atributos

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Num. De habitaciones	100.0%	12	Tiene terrazas	13.2%
2	Año de construcción	99.7%	13	El piso tiene anejos	11.9%
3	Número de inmuebles en finca	56.1%	14	Canal de venta del inmueble	8.9%
4	Calidad de la construcción	48.1%	15	Es un estudio	8.4%
5	Número de pisos en edificio	44.0%	16	Tiene armarios empotrados	7.6%
6	Planta en el edificio	24.9%	17	Posición vertical en el edificio	7.6%
7	Tiene ascensor	16.2%	18	¿Tiene su edificio una piscina?	7.1%
8	Tipo de instalaciones de la finca	15.6%	19	Es un dúplex	6.5%
9	¿Interior o exterior?	15.0%	20	Es un ático	4.1%
10	¿Tiene aire acondicionado?	14.2%	21	Nuevo o segunda mano	3.4%
11	Periodo en Año-Mes	13.8%	22	Tiene balcón	1.3%

Fuente: elaboración propia

Tabla 5.25. Importancia de variables plurifamiliar modelo de utilidad

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Superficie total construida	100.0%	15	Accesibilidad WALK 6	14.3%
2	Población del municipio	61.6%	16	Accesibilidad CAR 4	14.3%
3	Accesibilidad CAR 2	44.6%	17	Accesibilidad WALK 7	14.0%
4	Densidad de población	34.4%	18	Accesibilidad WALK 8	13.6%
5	Proporción alquiler/venta	33.4%	19	Accesibilidad WALK 9	12.9%
6	Num. de inmuebles en alquiler	25.6%	20	Pct. Estudios estudios superiores	12.7%
7	Accesibilidad WALK 1	24.5%	21	Accesibilidad WALK 3	12.4%
8	Número de contactos en la zona	23.1%	22	Accesibilidad CAR 9	11.9%
9	Accesibilidad WALK 2	20.9%	23	Accesibilidad CAR 3	11.6%
10	Accesibilidad WALK 4	20.1%	24	Accesibilidad CAR 6	11.3%
11	Accesibilidad CAR 5	19.3%	25	Accesibilidad CAR 7	10.9%
12	Accesibilidad CAR 8	18.1%	26	Porcentaje de mayores	10.7%
13	Accesibilidad CAR 1	15.5%	27	Tasa de extranjeros	9.8%
14	Accesibilidad WALK 5	14.4%	28	Inmuebles en venta en la zona	7.8%

Fuente: elaboración propia

El modelo utilidad combina los atributos constructivos básicos con los de localización y los de dinámicas del mercado (*LEADS*, *RENT_SALE_RATIO* y *LEADS_RESIDENTIAL*), como se aprecia en la Tabla 5.25. Todos atributos

accesibilidad mantienen una importancia similar, destacando ligeramente aquellas procedentes del medio de transporte a pie. Como sucede en las múltiples referencias de la literatura, no existe una gran diferencia entre la importancia de las distintas medidas de accesibilidad, repartiéndose la importancia casi de forma equitativa, con la tendencia de aparecer ciertas medidas demográficas como la densidad de población en las primeras posiciones, como en (Rico y Taltavull, 2021).

Tabla 5.26. Importancia de variables unifamiliar, modelo de atributos

Pos.	Variable	Importancia	Pos.	Variable	Importancia
1	Inmuebles en alquiler	100.0%	18	Canal de comercialización	7.8%
2	Superficie construida	93.5%	19	¿Tiene su edificio una piscina?	7.7%
3	Área útil	90.8%	20	Periodo Año-mes	7.6%
4	Inmuebles en venta	59.8%	21	Tiene armarios empotrados	6.5%
5	Proporción alquiler/compra	58.5%	22	¿Tiene aire acondicionado?	6.2%
6	Tipo de zona (cluster)	57.2%	23	Tiene jardín	6.0%
7	Contactos medios en zona	40.8%	24	Tamaño del garaje	5.6%
8	Año de construcción	30.1%	25	Tiene almacenamiento / Trastero	4.8%
9	Tamaño de la parcela	26.0%	26	Está orientado al sur	4.3%
10	Calidad de la construcción	22.9%	27	Tiene terrazas	4.0%
11	Número de habitaciones	20.4%	28	Está orientado al norte	3.6%
12	Número de dormitorios	20.3%	29	Tiene portero	3.6%
13	Número de baños	20.1%	30	Está orientado al este	3.1%
14	Número de pisos en edificio	10.9%	31	Nuevo o segunda mano	2.7%
15	Tipo de inmueble unifamiliar	10.5%	32	Está orientado al oeste	2.6%
16	Certificado energético	8.9%	33	Tipo de garaje	0.1%
17	Tipo de instalaciones en finca	8.6%			

Fuente: elaboración propia

Para el modelo de atributos de unifamiliar, los atributos de dinámicas de mercado junto con los constructivos y el clúster de zonas, acaparan la mayor parte de la reducción de la impureza, como se ver en la Tabla 5.26.

En el modelo de utilidad unifamiliar, como se aprecia en la Tabla 5.27, la importancia relativa de las medidas de accesibilidad es mucho mayor que en el caso de plurifamiliar. Por otra parte, la variable más relevante es la población del municipio (en lugar de la superficie).

Tabla 5.27. Importancia de variables plurifamiliar, modelo de utilidad

Pos.	Variable	Imp.	Pos.	Variable	Imp.
1	Población del municipio	100.0%	13	Accesibilidad WALK 7	46.2%
2	Accesibilidad CAR 2	84.1%	14	Accesibilidad CAR 3	46.1%
3	Accesibilidad CAR 5	73.0%	15	Accesibilidad WALK 2	45.7%
4	Accesibilidad WALK 6	70.4%	16	Accesibilidad CAR 9	45.3%
5	Pct. personas con estudios superiores	65.3%	17	Accesibilidad WALK 9	44.8%
6	Densidad de población	63.0%	18	Accesibilidad CAR 7	43.8%
7	Accesibilidad CAR 8	62.1%	19	Accesibilidad WALK 8	41.0%
8	Accesibilidad WALK 1	54.3%	20	Accesibilidad CAR 6	40.2%
9	Accesibilidad WALK 5	53.9%	21	Accesibilidad WALK 3	38.5%
10	Accesibilidad WALK 4	53.3%	22	Porcentaje de mayores	33.4%
11	Accesibilidad CAR 4	52.3%	23	Tasa de extranjeros	28.8%
12	Accesibilidad CAR 1	47.4%			

Fuente: elaboración propia

Estos resultados guardan consistencia con otros estudios[17], pero sería interesante desarrollar un análisis más en profundidad, ya que como demuestran empíricamente Rico y Taltavull (2021), la influencia de las variables difiere entre los estratos de la muestra.

El método propuesto para construir los de modelos hedónicos de oferta en el presente Capítulo, cuenta con un alto grado de desglose y permite introducir una mayor precisión a las estimaciones del modelo de mercado, presentado en el Capítulo 3. Sin embargo, la aplicación de estas estimaciones requieren un ajuste del planteamiento original del modelo de conversión oferta-mercado, debido a los sesgos derivados de la ausencia de control de la localización. En el próximo capítulo se presenta un método que corrige este efecto, y permite calcular estimaciones del precio de mercado con un alto grado de detalle.

En este capítulo se ha desarrollado un modelo que estima,de forma precisa los precios de oferta, y que aprovecha la contribución de varios enfoques para crear un estimador más robusto. En el siguiente capítulo, se presentará una metodología capaz de corregir los sesgos zonales de los que adolece el modelo hedónico de mercado presentado en el capítulo 3.

[17] Como por ejemplo, Rico y Taltavull (2021), Füss y Koller (2016) o Ĉeh (2018).

Anexo 5a. Clasificación automática de zonas

Como paso previo a la construcción del modelo ensamblado, se crea una variable sintética que pueda relacionar los distintos modelos atendiendo a la diversa naturaleza de las zonas. Este tipo de zona (denominado como *Cluster* en los modelos) que representa una categorización natural de secciones censales. Esta categorización se desarrollada mediante el método de aprendizaje automático no supervisado K Medoides (Schubert y Rousseeuw, 2019), que es una versión robusta del algoritmo de análisis clúster K Medias (Lloyd, 1982). Para evitar distorsiones de escala y colinealidad se preprocesan las variables mediante un proceso de escalado multimensional (SMACOF) a través de un matriz simétrica de disimilitud (Borg y Groenen, 2005).

Se trabaja en dos ámbitos: el primero para Madrid, y el segundo, para el resto de la Comunidad, con 7 grupos distintos en cada ámbito (K=7). Las fuentes de información son datos los de renta y viviendas de la Comunidad de Madrid [18], puntos de interés de Open Street Map (2017).

Para cada sección censal de la capital, se usan datos que recogen las características zonales de tipo demográfico, económico, y de oferta de servicios de turismo y ocio, que se reduce a las siguientes variables:

- Número de viviendas.
- Número de comercios.
- Número de bares-restaurantes.
- Número de hoteles.
- Número de museos.
- Número de monumentos.
- Número de negocios de tipo industria.
- Paradas de transporte público.
- Nivel educativo.
- Porcentaje de población extranjera.
- Número de centros sanitarios
- Renta media por persona.

Mientras que para el resto de provincia, se parte de la misma base, excluyendo las de tipo turístico:

- Número de viviendas.
- Número de comercios.
- Número de bares-restaurantes.
- Número de hoteles.

[18] Portal estadístico de la Comunidad de Madrid (https://www.madrid.org/iestadis).

- Número de negocios de tipo industria.
- Renta media por persona.

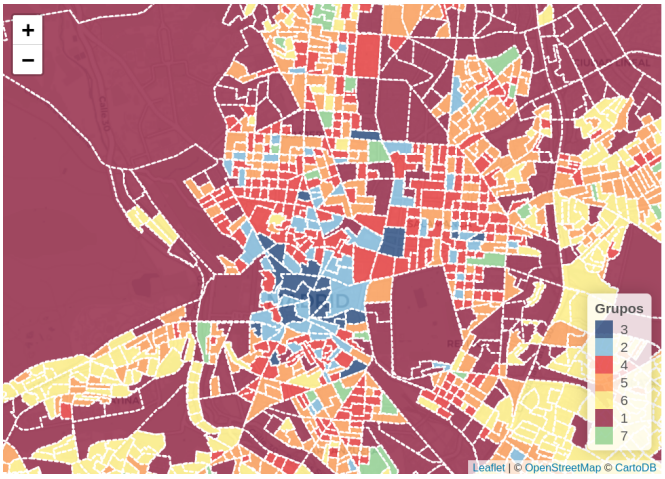
Se generan 7 grupos de zona, numerados del 1 al 7, y descritos en la Tabla 5.28.

Tabla 5.28. Descripción de grupos para la ciudad de Madrid

Código	Descripción	Secciones censales
1	Zona con la renta más alta. Muy baja densidad de viviendas, comercios y bares-restaurantes	471
2	Gran densidad de viviendas y lugares de ocio. Alta densidad de comercios y bares-restaurantes	66
3	Alta densidad de viviendas. Gran densidad de comercios, bares-restaurantes, hoteles y monumentos	21
4	Gran densidad de viviendas. Alta densidad de comercios y bares-restaurantes	313
5	Media densidad de viviendas, comercios y bares-restaurantes	587
6	Zona con la renta más baja. Baja densidad de viviendas, comercios y bares-restaurantes	897
7	Zona industrial. Baja densidad de viviendas. Media densidad de comercios y bares-restaurantes	51

Fuente: elaboración propia

Figura 5.11. Clusters de zona en el municipio de Madrid



Fuente: elaboración propia.

En el centro (véase Figura 5.11), dónde ubica la actividad turística y de ocio, se produce una alta concentración de los grupos 2 y 3; los grupos 4 y 5 se extienden desde el centro hacia el norte y el sur por el eje Prado-Castellana (división norte

sur de la ciudad); los grupos 6 y 1 se ubican en el anillo exterior, con mayor presencia del grupo 1 en el norte y oeste; y por último, el grupo 7 solamente está presente en el distrito de Villaverde (sur).

La agrupación zonal para el resto de la Comunidad[19], recogida en la Tabla 5.29, distingue principalmente las área rurales y las áreas residenciales de rentas altas. En las últimas, predomina la vivienda de tipo unifamiliar y se concentran en la parte norte-oeste del anillo de municipios inmediatamente exterior a la capital.

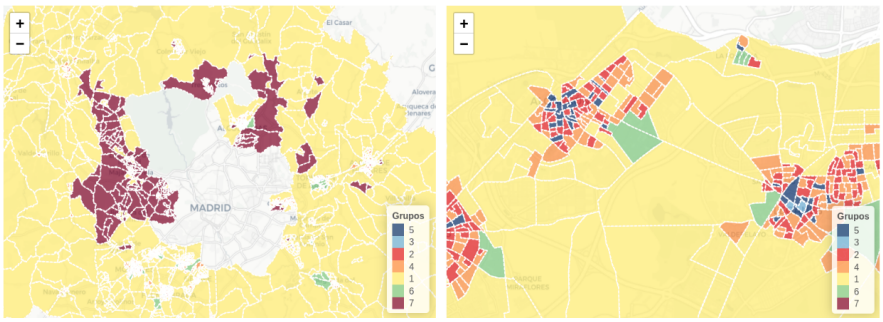
Tabla 5.29. Descripción de grupos resto de la Comunidad de Madrid

Código	Descripción	Secciones censales
1	Muy baja densidad de viviendas, comercios y bares-restaurantes	652
2	Alta densidad de viviendas, comercios y bares-restaurantes	371
3	Alta densidad de viviendas, comercios, bares-restaurantes y hoteles	18
4	Media densidad de viviendas, comercios y bares-restaurantes	510
5	Gran densidad de viviendas, comercios y bares-restaurantes	107
6	Zona industrial. Baja densidad de viviendas. Densidad de comercios y bares-restaurantes	30
7	Zona con la renta más alta. Baja densidad de viviendas, comercios y bares-restaurantes	175

Fuente: elaboración propia

Estas últimas agrupaciones tienen un menor nivel de variabilidad, como se observa la Figura 5.12.

Figura 5.12. Cluster resto de la Comunidad de Madrid



Fuente: elaboración propia.

[19]Aunque las zonas compartan los códigos numéricos del clustering de la ciudad se refieren a tipos de zonas distintas.