Capítulo 6

Modelo hedónico final

“El conocimiento no sirve para nada a menos que se ponga en práctica”

— Anton Chejov

6.1 Introducción

Los capítulos anteriores se han centrado en la construcción de un modelo preciso que represente de forma fiel el comportamiento del mercado, que replique los valores de las estadísticas oficiales disponibles. Asimismo, se estima una función que transforma los precios de oferta en precios reales de alquiler. El modelo de oferta, por otra parte, permite estimar los precios en comercialización según un conjunto muy extenso de características, aportando herramientas precisas de estudio de los factores que contribuyen al valor de mercado de las viviendas.

El método anterior produce series temporales de las rentas con un alto grado de descomposición, reduciendo los sesgos de omisión de variables de oferta que existen con el uso de los datos originales de fuentes oficiales. Sin embargo, existe aún un inconveniente a resolver en esta aproximación, identificado en el apartado 3.4.3 del Capítulo 3, que es la dificultad del modelo de mercado de reproducir los comportamientos específicos de cada zona. Esta cuestión está motivada, principalmente, por la ausencia de información geográfica específica en los datos de la EPF principalmente, y para solucionarlo requiere de una corrección zonal de precios.

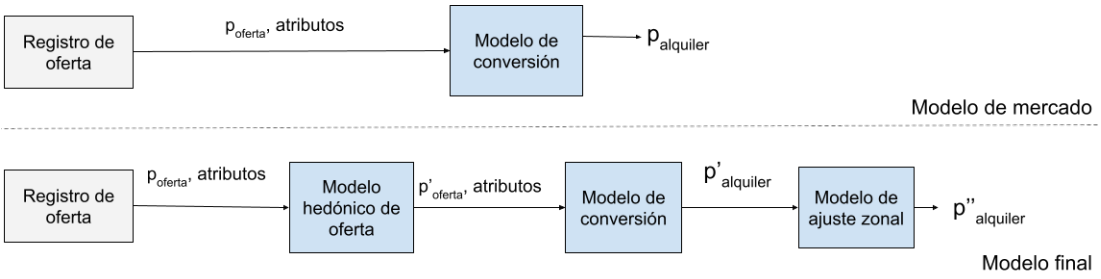
Este capítulo presenta un modelo hedónico denominado “final”, cuyo objetivo es corregir los desequilibrios zonales de los métodos usados en los capítulos anteriores. Para ello, se utilizan las series de precios anuales de alquiler del “Sistema Estatal de Índices de Referencia del Precio del Alquiler de Vivienda”,

publicadas por el MITMA (2020), y cuya utilidad se apoya en una alta correlación entre estos datos y los de oferta (véase (Rey-Blanco *et al.*, 2023a)), tanto en sus valores absolutos como en sus variaciones interanuales.

La cointegración entre el dato de oferta y el final ha sido ampliamente analizada en la literatura, y se deriva del proceso secuencial en el que se forman los distintos precios (Shimizu *et al.*, 2016). De igual manera, significar estudios como el de Kokot (2015), sobre precios de alquiler en Polonia, o el de Ardila *et al.* (2021) para el mercado suizo.

El proceso de cálculo del modelo final se realiza en los tres pasos descritos en la Figura 6.1. El flujo superior, muestra el cálculo del precio de alquiler por el modelo de mercado original, $p_{alquiler}$; la secuencia de la fila inferior, representa el proceso completo del modelo final, a través del encadenado de tres modelos: el hedónico de oferta, el modelo de conversión a precios de mercado y el modelo de ajuste zonal, con el resultado final $p''_{alquiler}$.

Figura 6.1. Proceso de estimación del precio del alquiler usando precio de oferta del modelo hedónico



Fuente: elaboración propia.

El capítulo se estructura en dos partes: la primera, contiene una descripción metodológica del proceso de corrección de sesgo zonal, y la segunda, que analiza los resultados obtenidos desde las ópticas de reducción del sesgo zonal, sesgo general de los precios, y validez del modelo para proyectarse a periodos futuros.

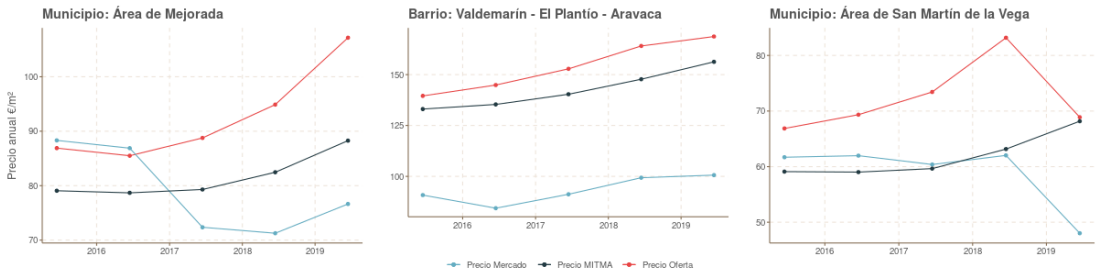
6.2 Metodología

El sesgo identificado en el apartado 3.4.3 del Capítulo 3, procede de la incapacidad de controlar la heterogeneidad espacial en los precios de la vivienda con los atributos de los datos de la EPF. Esta heterogeneidad, significa que la lógica sobre que se basa la formación de los precios no es un proceso estacionario espacialmente, por tanto, la relación funcional entre las covariables y el precio tiene peculiaridades diferentes en función de la unidad geográfica de análisis. En el Anexo I del presente capítulo, se describe en mayor profundidad las cuestiones de identificación y corrección de sesgos aplicadas en la metodología.

El fenómeno de la heterogeneidad espacial está ampliamente documentado y se analiza con profundidad por Hu (2022), Wu (2020), Helbich (2014), Páez (2008) y Kestens (2006). En nuestro caso, sus consecuencias se resumen en que el modelo, al no disponer de referencia a las zonas concretas, tiende a representar los precios de las zonas como un compuesto de las medias del valor de un estrato macro zonal (por ejemplo: zona periférica de altos ingresos o municipio de más de 50.000 habitantes), de lo que se deriva la imposibilidad de controlar la variabilidad de precios, como se verá más adelante en los ejemplos de las Figuras 6.2.

La ausencia de control por la zona geográfica provoca una asignación de precios arbitraria ante un desglose de datos geográfico-funcional, puesto que el modelo solo es capaz de asignar precios a través de la dimensión funcional. Como consecuencia, se producen dos fenómenos: 1) el dato desglosado por zona muestra una alta irregularidad en el tiempo[1], y 2) al calcular agregados con más de una dimensión de estratificación, el resultado también muestra irregularidades.

Figura 6.2. Precios de mercado, oferta y precios MITMA, vivienda plurifamiliar



Fuente: elaboración propia.

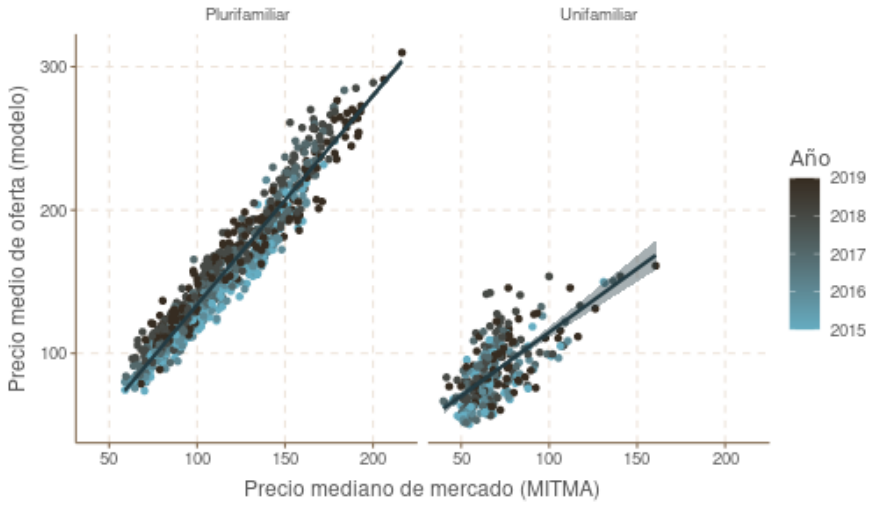
Para demostrar esta situación gráficamente y a modo de ejemplo, se representan en la Figura 6.2, los precios de mercado, oferta y de MITMA para tres zonas

[1]El término "irregularidad" se refiere a secuencias de precios incoherentes con respecto a la tendencia general del mercado, por ejemplo inconsistencias en la evolución del precio de mercado con respecto al precio de oferta.

diferenciadas. Para el caso de Mejorada del Campo[2], se observa que la serie de oferta y la de MITMA están cointegradas, mientras que la serie del modelo de mercado tiene una tendencia totalmente distinta a partir de 2016. En el caso de El Plantío-Aravaca[3], se aprecian diferencias en las pendientes de las series y una gran diferencia entre el precio medio del modelo y el registrado por MITMA . El tercer ejemplo, para San Martín de la Vega[4] muestra una anomalía en la serie de oferta y mercado para el año 2019, atribuible a anomalías del precio de ofert a.

Una propiedad muy interesante de estas series de precios es la alta correlación entre los precios de mercado y de oferta. A este respecto, existen varios artículos que analizan el nivel de relación entre las poblaciones de oferta y de transacción: Chapelle *et al.* (2022), para Francia, y Kokot (2015), en Polonia, estudian el nivel de correlación de transacciones y oferta obteniendo un coeficiente de correlación de 0,95 y 0,99 respectivamente.

Figura 6.3. Correlación entre precio medio de oferta y mediano del MITMA



Fuente: elaboración propia.

La correlación entre precios es mucho más fuerte en viviendas plurifamiliares que en unifamiliares, como se aprecia en la Figura 6.3. Para las últimas, la variabilidad aumenta a medida que se incrementan los precios del mercado. Por otra parte, se observa una relación cambiante en el tiempo, atribuible a las condiciones coyunturales del mercado, por ejemplo, la pendiente menos pronunciada del 2015 puede indicar una fase de mayor rotación de los alquileres. Existen distintas referencias bibliográficas que demuestran que la relación precios de oferta/mercado no son inmutables, como la aportación de Han (2016),

[2] Se toma este municipio por ser un área con una muestra pequeña e irregular.
[3] Se toma la zona por estar en el área metropolitana cercana a Madrid y tener una amplia muestra.
[4] Se selecciona esta zona por ser un área rural lejana a la capital.

quien muestra como se produjeron variaciones de entre el 15 y 30%, durante el boom inmobiliario de los años 2000-2007 en Estados Unidos.

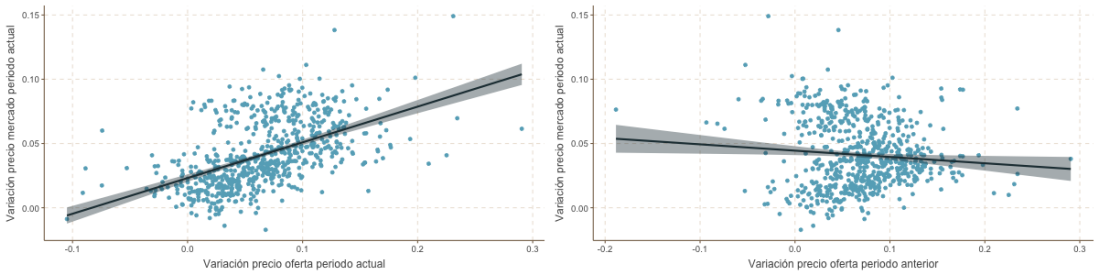
Las variabilidad de la Figura 6.3 se puede, además, justificar por la diferente de composición de ambas poblaciones[5], aún cuando los valores medios de las poblaciones estén fuertemente correladas. A este respecto, Shimizu (2016), Kolbe (2021) y Ardila *et al.* (2021) han constatado que las distribuciones de frecuencias de precios oferta y de mercado no son iguales, y cuyas diferencias se pueden deber a una sobrerrepresentación en oferta de los inmuebles menos líquidos y la infrarrepresentación de los más líquidos, o también, a la presencia de anuncios en régimen de subasta (Han y Strange, 2016). Por otra parte, Ardila *et al.* (2021) destacan que el dato de oferta tiende a infraestimar la magnitud de los cambios de tendencia.

Por otra parte, Diaz y Jerez (2013) argumentan que las fricciones en el proceso de búsqueda produce retrasos en dicho proceso y, consecuentemente, introduce volatilidad en los precios. Estas fricciones están presentes en manera desigual en los submercados que componen la muestra por los desequilibrios entre la oferta o la demanda, por lo que la única forma de controlarlas en los modelos es mediante el uso de variables que representen las características de cada mercado.

La cointegración entre series de alquiler y oferta también ha sido documentada en la literatura. Por ejemplo, Kokot (2015) comprueba la existencia de cointegración más un retraso temporal de 5 meses entre la serie de mercado y la de oferta. En nuestro caso, se han analizado la relación entre las variaciones anuales del precio de alquiler y las de oferta, identificando que las variaciones en los precios de mercado están correlacionadas con la variación del precio de oferta del año anterior, como se puede comprobar gráficamente en la Figura 6.4. Esta relación no es tan fuerte como la que existe entre los precios, porque en este caso el coeficiente de correlación de Pearson entre las variaciones es de 0,51. Este menor grado de relación podría estar motivada por dos cuestiones: 1) que el periodo de retraso entre series sea menor de un año, como el caso de (Kokot y Bas, 2015), y 2) que esta relación pueda variar en función de la zona. El segundo argumento se sustenta sobre la dependencia de la ratio de precios y la intensidad de la demanda de la zona (Han y Strange, 2016; Han y Strange, 2014; Shimizu *et al.*, 2016). Por tanto, como el modelo no incorpora estos efectos en un mercado heterogéneo, es plausible la existencia de variabilidad no controlada por este motivo.

[5]En el caso de la población de MITMA se desconocen.

Figura 6.4. Relación variación precio de mercado y oferta (vivienda plurifamiliar)



Fuente: elaboración propia.

Sobre la base anterior, se construyen dos modelos de regresión por mínimos cuadrados ordinarios, uno para cada tipologías de vivienda. Las variables independientes son la variación de oferta con uno y dos periodos de retraso. Para el caso de el tipo plurifamiliar, los coeficientes del modelo se recogen en la Tabla 6.1, y se observa como las variaciones de precios son altamente representativas. Los modelos permiten reconstruir la variación del precio de mercado a partir del histórico de variaciones del precio de oferta. Lo que desde un punto de vista económico representa la capacidad de la zona para absorber los inmuebles en oferta, de tal forma que se sustituyen inmuebles en oferta por inmuebles de mercado, produciéndose una transferencia del precio de oferta sobre el de mercado.

Tabla 6.1. Coeficientes del modelo de absorción - viviendas plurifamiliares

	Estimación	std.error	t value	Pr(>|t|)	signif.
(Intercept)	0.03	0.00	16.12	< 2e-16	***
VAR_ASKING	-0.07	0.02	-4.57	6.10e-06	***
VAR_ASKING_1	0.30	0.02	19.55	< 2e-16	***

Códigos signif.: *** 0,001 ** 0,01 * 0,05 . 0,1 1
Error estándar de los residuos: 0.0169 sobre 564 grados de libertad (DF)
R^2: 0,423, R^2 ajustado: 0,421
F-statistic: 2017 sobre 2 y 564 DF, p-value: < 2,2e-16
Num. observaciones: 564

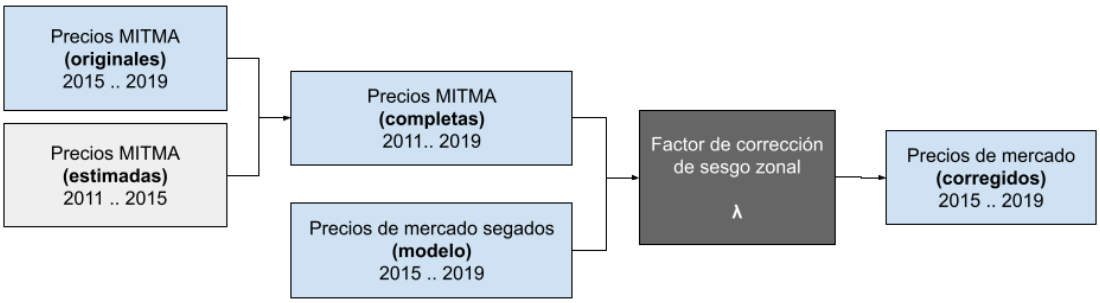
El coeficiente de determinación de la regresión anterior indica que el modelo no explica totalmente la varianza de la variable objetivo. Es evidente que parte de esta varianza es explicada mediante la introducción de nuevas variables, como el nivel de demanda de la zona o la evolución de volumen inmuebles en oferta (De Wit *et al.*, 2013). No obstante, el uso de series de precios anuales provoca que la varianza debida a retrasos menores a 12 meses sean difícilmente controlables.

La metodología se basa en la estimación de una serie de ratios simples[6] sobre un una muestra estratificada (Lohr, 2019), aplicada en dos etapas. El proceso crea, para cada estrato e (definido como la combinación entre zona, tipo de vivienda y un año t), un factor dinámico de corrección de sesgo del precio, denominado en adelante $\lambda_{t,e}$, y que se aplica para calcular los nuevos precios de mercado $\hat{P}$:

$$\hat{P}_{t,e} = \lambda_{t,e} \times P_{t,e} \quad [6.1]$$

El proceso se describe, de forma general, en la Figura 6.5.

Figura 6.5. Pasos del proceso de corrección del sesgo zonal



Fuente: elaboración propia.

Los factores a aplicar a los precios se calculan sobre los datos MITMA disponibles[7]. Los factores λ relacionan el precio medio ponderado con la mediana del precio, para el estrato de la muestra conocida, P^M. Este factor de ajuste, denominado $\lambda^M_{t,e}$, se calcula a través de la expresión:

$$\lambda^M_{t,e} = \frac{P_{t,e}}{P^M_{t,e}}, \forall t \in [2015..2019] \quad [6.2]$$

donde t representa a un año entre 2015 y 2019.

Para los periodos anteriores, en los que no se dispone dato de MITMA, se utiliza un promedio entre el valor original de mercado y el precio estimado por el modelo de regresión de la Tabla 6.1. Calculado según la expresión:

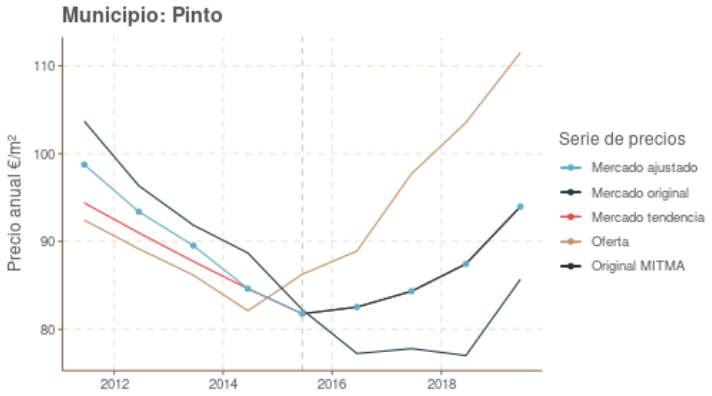
$$\lambda^M_{t,e} = \frac{P_{t,e}}{\omega \cdot \hat{P}^V_{t,e} + (1-\omega) \cdot (P_{t,e} \times \lambda_{2015,e})}, \forall t \in [2011..2014] \quad [6.3]$$

[6] Los ratios de ajuste del muestreo se han denominado factores de corrección dinámicos o λ.
[7] Las series de datos MITMA solo están disponibles a partir de 2015.

donde $\hat{P}^V_{t,e}$ es el precio del modelo de mercado original, y $P_{t,e} \times \lambda^M_{2015,e}$ el precio MITMA estimado a pasado, entre 2011 y 2015. El factor ω indica la proporción de la tendencia que se toma del modelo original, que en este caso se utiliza un valor de $\omega = 0,5$, y que está sujeto a revisión en futuras actualizaciones de la metodología.

El modelo final corrige el sesgo para cada macro-estrato zonal, definido como tipo (unifamiliar, plurifamiliar y zona). Por tanto, la serie final con el ajuste de sesgo se obtiene multiplicando el precio individual por la ratio de corrección λ del estrato.

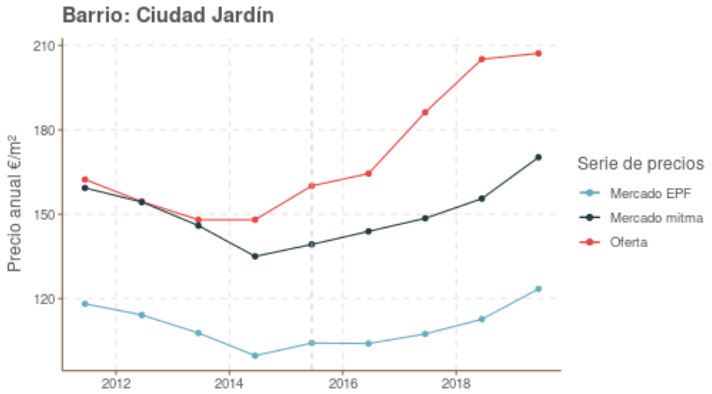
Figura 6.6. Series de precios de mercado ajustada



Fuente: elaboración propia.

Como se observa en la Figura 6.6, los precios anteriores al 2015 se estiman a partir de los precios del modelo corregidos con la regresión de absorción de la oferta, y los posteriores, son ajustados mediante el dato MITMA de la zona. Para el caso de Pinto[8], en la tipología plurifamiliar, se aprecia la corrección de la tendencia en la serie de precios del mercado sobre el periodo 2016-2018.

Figura 6.7. Niveles de precios de series de mercado: EPF y MITMA



Fuente: elaboración propia.

[8]Se toma este ejemplo porque la serie de alquiler muestra un comportamiento incoherente con MITMA y oferta para el periodo entre 2016 y 2019.

Sin embargo, el precio de las rentas de las series de MITMA no se utiliza la misma magnitud que en la encuesta de la EPF, la cual contiene alquileres no sujetos a las declaraciones del IRPF, como son alquileres sociales o imputados. Se puede comprobar de manera gráfica sobre la Figura 6.7, la diferencia de escala entre las series de precios para el barrio Ciudad Jardín en Madrid[9].

Para efectuar el ajuste final a los niveles de precios de alquiler para la EPF, se calcula una tasa de ajuste final $\lambda_{t,e}$ a través de unos ponderadores w_e para cada estrato e, calculados como la proporción histórica entre el nivel estimado de precios MITMA y el precio de las series originales basadas en la EPF. El ponderador realiza un re-escalado de las series de precios de mercado, para que se ajusten a las medias de las series de la EPF desglosadas funcionalmente, y se elimine el sesgo de escala. El cálculo se realiza, por tanto, mediante la expresión siguiente:

$$\lambda_{t,e} = w_e \times \lambda_{t,e}^{M} \qquad [6.4]$$

Y por consiguiente, el precio final ajustado $\hat{P}_{t,e}$ se define como:

$$\hat{P}_{t,e} = \lambda_{t,e} \times P_{t,e} \qquad [6.5]$$

De forma alternativa, la expresión anterior se podría haber expresado en función del producto entre el factor MITMA (λ^M) y por la proporción de los pesos de la EPF (w_e), es decir:

$$\hat{P}_{t,e} = w_e \times \lambda_{t,e}^{M} \times P_{t,e} \qquad [6.6]$$

El análisis de los valores λ^M tiene una utilidad adicional, permite medir el grado de sesgo incurrido en el modelo de precios original.

[9] Se toma una zona urbana con buen nivel de muestra, en la que se aprecia notablemente la diferencia de escala entre el precio MITMA y el de alquiler.

6.3 Resultados

Para evaluar los objetivos perseguidos de incorporar frecuencia mensual a las series, y corregir la coherencia geográfica de los precios de mercado, se han analizado los resultados del modelo final en torno a cuatro cuestiones:

- El método es capaz de reducir el sesgo zonal del modelo de mercado original.
- Dado que los modelos de *bagging* son propensos a los sesgos, se evalúa su presencia en los precios finales de oferta y de mercado.
- Control la coherencia mensual y anual de las series de precios.
- Confirmar la capacidad de generalización en el tiempo del modelo de mercado.

6.3.1 Eficacia en el control del sesgo zonal

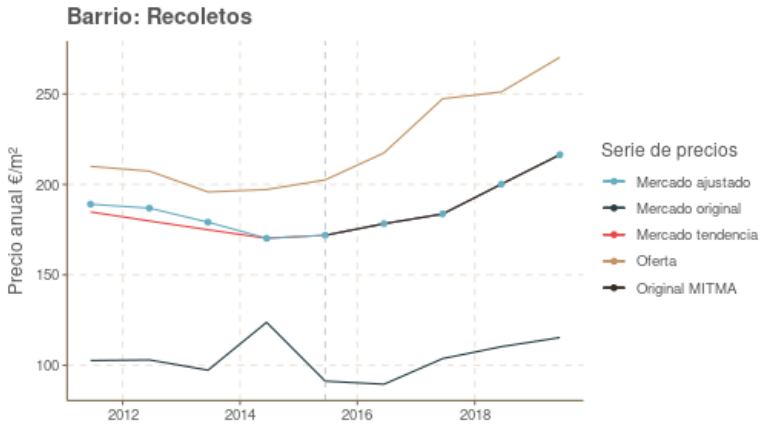
En esta sección se abordarán los dos sesgos principales del modelo, y que son la fuente principal de las distorsiones en los precios del modelo de alquiler:

- Sesgos debidos a la composición, y que se producen por que la calibración entre las poblaciones de oferta y alquiler no tiene encuentra los niveles de precios de las zonas.
- Sesgos propios de los precios de la oferta, que se manifiestan principalmente en zonas con menor nivel de muestra, donde hay una gran variabilidad del precio por unidad de superficie o la estimación está más expuesta a variables omitidas.

Para ilustrar de forma gráfica el método de ajuste, se representa a modo de ejemplo el resultado para del barrio de Recoletos en Madrid en la Figura 6.8 [10] (para ver todos los ejemplos, consúltese https://github.com/davidreyblanco/idx/tree/master/hedonic-final/unbias-fixed y los enlaces del Anexo I). Se observa un descenso anormal en el precio de mercado para el año 2015, sin una razón de mercado plausible que pueda justificarlo, y que es incoherente con los precios medianos del MITMA. La serie del modelo “final” propuesto, representada en la imagen como “Mercado ajustado”, corrige eficazmente la tendencia y ofrece una serie de precios de mercado coherente con el resto de series originales (MITMA y oferta).

[10] Se toma el caso de Recoletos para evidenciar el primero de los problemas (composición zonal), ya que cuenta con una muestra de anuncios numerosa y muestra una serie de precios de alquiler incoherente con los precios de oferta y los registrados en MITMA, particularmente para el año 2014.

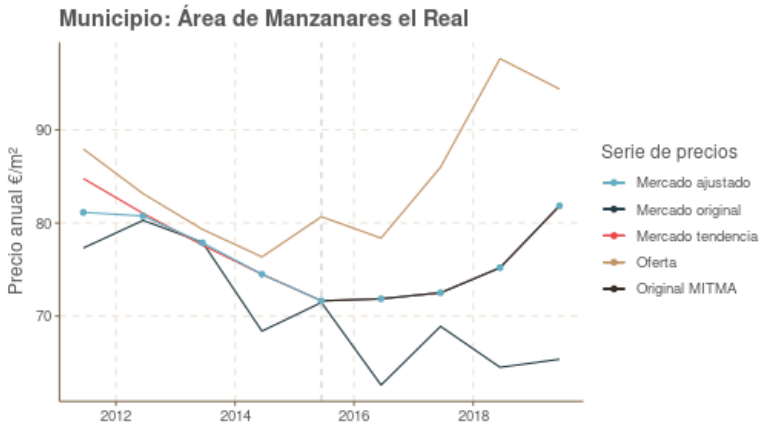
Figura 6.8. Series de precios vivienda plurifamiliar: barrio de Recoletos



Fuente: elaboración propia.

Es habitual que las zonas con muestras más pequeñas e irregulares den lugar a series de precios con variabilida espuria como podemos ver en Kokot y Bas (2015) o Eurostat (2013). Este fenómeno también se aprecia en nuestro caso, por ejemplo, con los datos para el municipio de la sierra madrileña de Manzanares el Real (municipio rural con un mercado inmobiliario poco activo), mostrados en la Figura 6.9. Se observa como la serie corregida controla las anomalías de modelo de mercado hasta 2017, y la excesiva caída de precios medios en oferta de 2019.

Figura 6.9. Series de precios vivienda plurifamiliar: Manzanares el Real

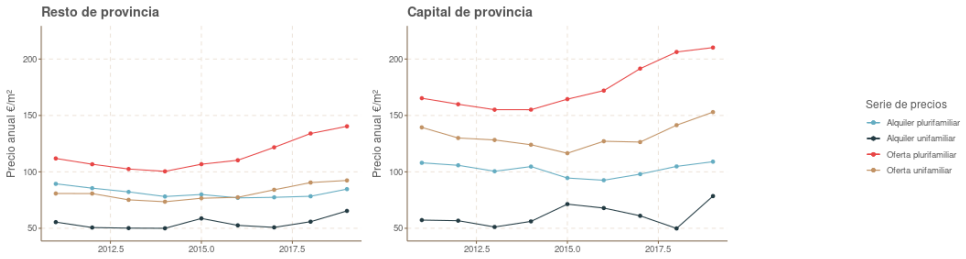


Fuente: elaboración propia.

De la misma manera, se corrige la incidencia del efecto de composición. Las Figuras 6.10 y 6.11 muestran los precios agregados originales y corregidos, en las que se ajustan eficazmente los efectos mencionados en los párrafos anteriores. Debido a su mayor representatividad, la corrección ligeramente mejor para los

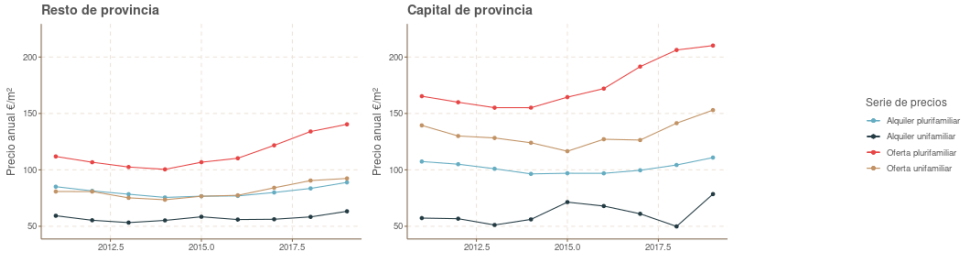
precios para viviendas plurifamiliares que en las unifamiliares.

Figura 6.10. Series de precios agregados, originales



Fuente: elaboración propia.

Figura 6.11. Series de precios agregados, corregidas



Fuente: elaboración propia.

Desde un punto de vista numérico, la Tabla 6.2 muestra como los precios de mercado ajustados por el modelo reducen de forma significativamente la variabilidad de los precios originales. Se aprecia, además, que se preservan los parámetros principales como la media y cortes cuantílicos.

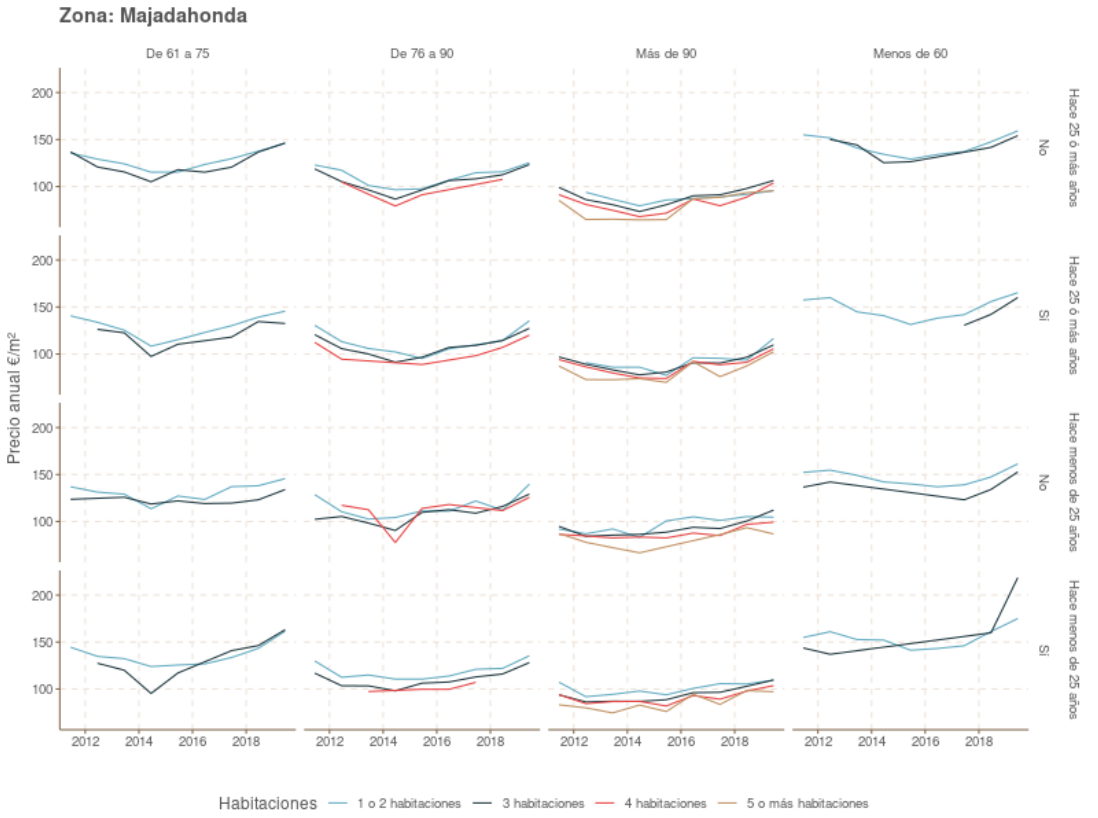
Tabla 6.2. Comparativa de agregados de series antes y después del ajuste zonal

Capital	Tipo	Paso	Media	Desv	Min	p25	p50	p75	Máx.
No	Plurifamiliar	Ajustado	80.73	4.44	75.59	76.91	79.97	83.54	88.94
		Original	81.43	4.33	76.97	78.16	79.96	84.72	89.41
	Unifamiliar	Ajustado	57.27	2.95	53.27	55.34	56.23	58.41	63.27
		Original	54.40	5.13	49.99	50.67	52.60	55.84	65.40
Sí	Plurifamiliar	Ajustado	102.08	5.13	96.46	97.01	100.99	105.03	110.90
		Original	102.02	5.90	92.58	98.03	104.70	105.84	109.04
	Unifamiliar	Ajustado	61.12	9.62	49.88	56.12	57.27	67.94	78.54
		Original	61.12	9.62	49.88	56.12	57.27	67.94	78.54

Se verifica gráficamente la preservación de la coherencia al desagregar las series según criterios funcionales. A modo de ejemplo se toma el municipio

de Majadahonda, representativo de las zonas residenciales del extrarradio de Madrid, cuyos precios originales (Figura 6.12) muestran una coherencia temporal consistente con las series corregidas de la Figura 6.11. Se observa cierto grado de cointegración entre series y una mayor variabilidad en los segmentos de mayor superficie, como aquellos con 4 y 5 o más habitaciones, debido a una menor representatividad estadística en estos segmentos.

Figura 6.12. Precios de mercado ajustados por número de habitaciones, superficie, antigüedad y si dispone de piscinas



Fuente: elaboración propia.

Finalmente, se estudia de forma cuantitativa el sesgo del resultado del ajuste zonal, para lo que se definen las métricas de $Bias_{zonal,s}$ y $Bias_{funcional,e}$, que representan el nivel de divergencia de los precios.

El primero, referido al sesgo zonal de cada zona s, y calculado como el nivel de discrepancia de las variaciones de la serie de mercado con respecto a la serie MITMA, pretende medir el nivel de diferencia en la tendencia de ambas series para las distintos estratos zonales definidos por: zona, tipo de vivienda y año. El cálculo de discrepancia se realiza según la siguiente expresión:

$$Bias_{zonal,s} = \frac{\sum_{t=2019}^{2015} w_{t,s} \cdot \left(\Delta \hat{P}_{t,t-1,s} - \Delta P^{M}_{t,t-1,s}\right)}{\sum_{t=2019}^{2015} w_s} \text{ , } \forall t \in [2015..2019], s \in S \qquad [6.7]$$

donde $\Delta \hat{P}_{t,t-1,s}$ representa la variación logarítmica anual del precio de mercado para la zona s en un año t, $\Delta P^{M}_{t,t-1,s}$ es la correspondencia variación logarítmica de los precios medianos de MITMA y $w_{t,s}$ el peso poblacional de este estrato zonal-temporal.

El sesgo funcional $Bias_{funcional,e}$, representa las diferencias de precios medios ponderados de los datos generados por el modelo ($\overline{P_e}$) con respecto a los precios medios ponderados del resultado del ajuste zonal ($\overline{\hat{P}_e}$). El estrato funcional e se corresponde con la parte de la población de la celda de unidad mínima utilizada en la calibración, es decir, la combinación de las siguientes variables[11]: *CAPROV, TIPOCASA, NHABIT, TAMAMU, ANNOCON, TIPOEDIF, ZONARES, DENSI, HASBOXROOM, HASPARKINGSPACE, HASSWIMMINGPOOL, HASAIRCONDITIONING, SUTC y SUPERF*. La fórmula de cálculo de la métrica de sesgo funcional es la siguiente:

$$Bias_{funcional,e} = \overline{\hat{P}_e} - \overline{P_e} \qquad [6.8]$$

La Tabla 6.4 y la Tabla 6.3 muestran los descriptivos de las medidas de sesgo zonal y funcional respectivamente. Se observa que el sesgo funcional medio es prácticamente cero para viviendas plurifamiliares, las cuales representan más del 95% de la población. Solo en el caso de la vivienda unifamiliar fuera de la capital los valores del nuevo modelo son ligeramente superiores (2,66), y en el caso de Madrid es 0,0 porque no se realiza un ajuste zonal. Por tanto, se confirma la eficacia del método propuesto para eliminar el sesgo zonal, al ser prácticamente cero y con un nulo nivel de variabilidad.

[11] Véase el epígrafe 3.2.2.1, en el Capítulo 3, para más información sobre el significado y mayor información de las variables.

Tabla 6.3. Sesgo funcional después de ajuste funcional

Es capital	Tipo	N. estratos	Peso	Bias	Varianza bias
No	Plurifamiliar	90.705	43,1%	-0.27	90.16
	Unifamiliar	27.713	4,1%	2.66	299.67
Sí	Plurifamiliar	130.693	52,4%	0.00	175.36
	Unifamiliar	7.239	0,4%	0.00	0.00

Fuente: elaboración propia

Tabla 6.4. Sesgo zonal después de ajuste zonal

Es capital	Tipo	N. estratos	Peso	Bias	Varianza bias
No	Plurifamiliar	1.300	46,3%	0.01	0
	Unifamiliar	1.300	4,7%	0.01	0
Sí	Plurifamiliar	2.240	48,7%	0.01	0
	Unifamiliar	1.635	0,3%	0.00	0

Fuente: elaboración propia

6.3.2 Sesgos en los precios finales de oferta y mercado

Existen dos posibles fuentes de sesgo en los precios de oferta:

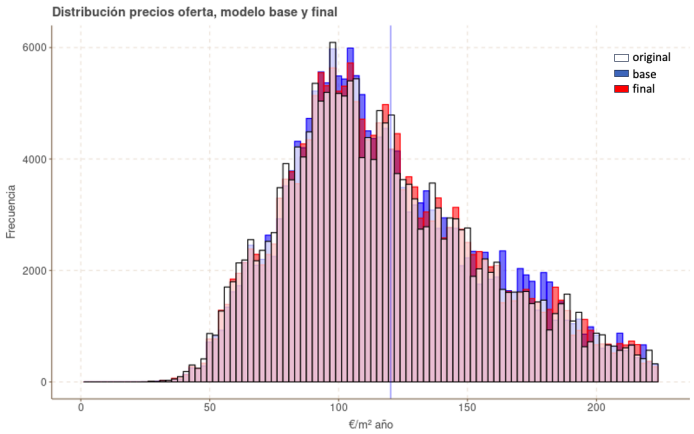
- Las diferencias en la distribución de valores de mercado y de oferta (Kolbe *et al.*, 2021; Ohnishi *et al.*, 2011; Shimizu *et al.*, 2016).
- Los estructurales de los modelos de tipo *Random Forests* (Hastie *et al.*, 2017), propensos a excluir los valores más extremos (Antipov y Pokryshevskaya, 2012), cuestión descrita en profundidad en el Anexo 6a del presente capítulo.

Al comparar los histogramas de frecuencias de precios ponderados[12] (Figura 6.13) no se observan diferencias significativas entre las tres valores: los originales (los presentes en los registros de originales de idealista), los del modelo de oferta del modelo de correspondencia (denominados “base” en la gráfica) y los finales (calculados por el modelo final de oferta). No obstante, se aprecia ligeramente una mayor concentración de los valores de los modelos alrededor del centro de masas, que coincide la propuesta de Antipov y Pokryshevskaya (2012), sobre que estos métodos tienden a eliminar los valores más extremos en los modelos, y por tanto a ofrecer peores estimaciones en estos casos[13] .

[12] Los valores de la representación se encuentran ponderados según sus pesos muestrales.

[13] Lo que no significa que los modelos funcionen incorrectamente, sino que el modelo no considera que son observaciones asociadas a comportamientos generalizables.

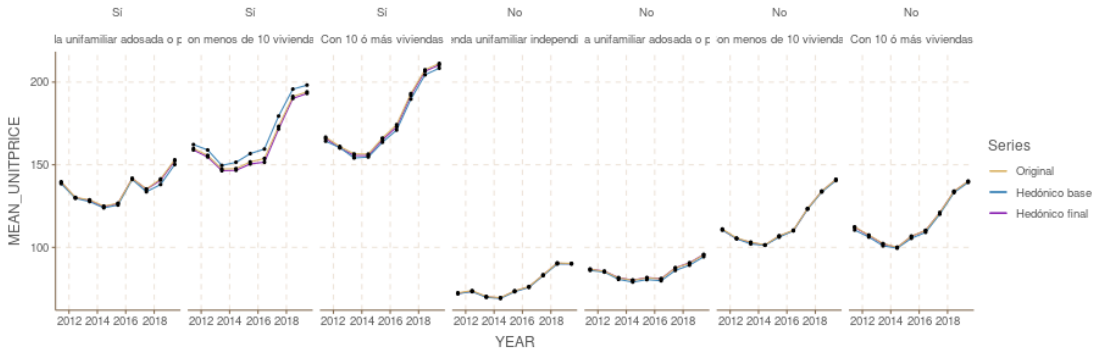
Figura 6.13. Distribución de precios de oferta



Fuente: elaboración propia.

Cuando se ponen en común las series temporales de precios de oferta (Figura 6.14), para los precios desglosados por capital de provincia y por tipo de edificio, se observa que los modelos base y final mantienen la tendencia de la serie de precios de oferta original. Los valores de las series tampoco muestran discrepancias en valores, exceptuando el caso de Madrid para edificios con menos de 10 viviendas, en el que el modelo final ajusta el sesgo del modelo base.

Figura 6.14. Precios de oferta desglosado por capital de provincia y tipo de edificio



Fuente: elaboración propia.

Para realizar el análisis en profundidad, y como no se dispone una correspondencia a nivel de registros observados y estimados[14], el sesgo se estima como el agregado de las contribuciones ponderadas de los sesgos de los estratos. En este caso, se estima de la forma siguiente:

[14] El valor observado de alquiler correcto sería el precio por el que se ha alquilado el inmueble en oferta.

$$Bias_Y = \sum_{e \in E} w_e \cdot Bias_Y(e) \quad [6.9]$$

donde $Bias_Y(e)$ es el sesgo del precio de oferta para el estrato funcional e[15], ponderado según el factor de elevación poblacional w_e del estrato.

Las medidas de sesgo de ambos modelos (Tabla 6.5) indican que tienden a infravalorar ligeramente, aunque el modelo hedónico final reduce a una tercera parte el sesgo del modelo original. Lo que indica que el modelo final es especialmente restrictivo con los valores extremos superiores.

El fenómeno de sesgo negativo en oferta es también identificado por Kolbe (2021), que lo atribuye a que los precios implícitos[16] son más dominantes en los modelos de oferta que en los modelos sobre operaciones reales.

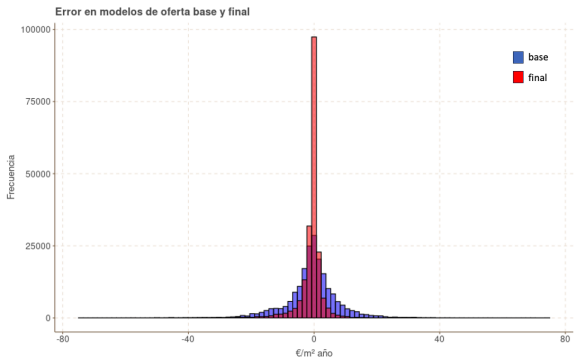
Tabla 6.5. Sesgo modelos hedónicos de oferta

Model	Media	Mediana	Pct. Media	Pct. Mediana	Desviación
Hedónico base	-0.97	-1.44	-0,67%	-0,99%	7.94
Hedónico final	0.57	0.14	0,39%	0,10%	3.36

Fuente: elaboración propia

De forma gráfica, se confirma el efecto de reducción de los errores en el modelo de oferta final 6.15. E indica cómo el uso de más variables afecta la variabilidad de los errores[17], a pesar de que el sesgo de los errores de ambos modelos sea muy cercano a cero.

Figura 6.15. Distribución de sesgos en modelos de oferta



Fuente: elaboración propia.

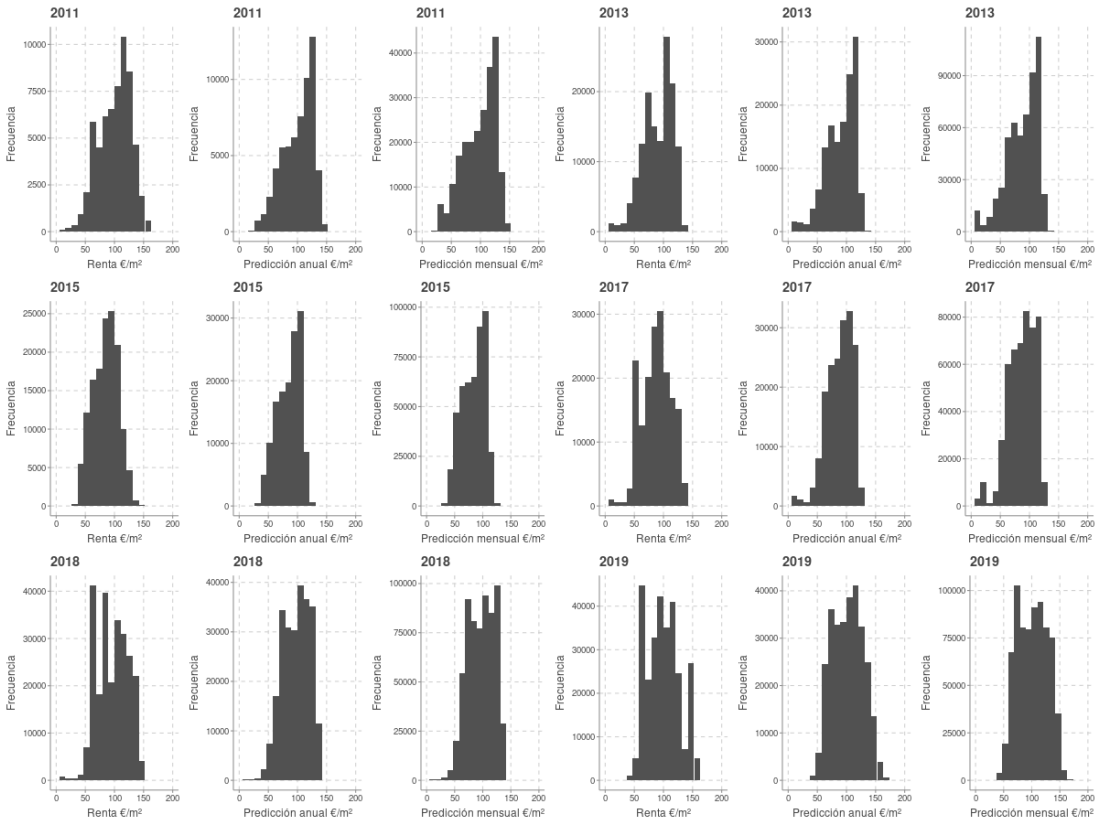
[15] Se sigue la misma estratificación funcional del apartado 6.3.1

[16] Los precios implícitos se refieren al valor económico de los factores no incluidos en el modelo, como son estado de conservación o cualquier otra variable omitida.

[17] Lógicamente, cuantas más variables estén disponibles menor será el riesgo de incurrir en sesgos de omisión de variables.

Para el caso de los precios de mercado, en la Figura 6.16 se comparan los histogramas de frecuencias para los precios de la EPF y los modelos, usando los elevadores muestrales de la calibración. Se observa que mientras que los precios de los modelos anuales y mensuales son equivalentes, los precios de la EPF son ligeramente diferentes. Por otra parte, la distribución de valores en cada conjunto varía en el tiempo, con un histograma más equilibrado en los últimos años de la serie, que se puede relacionar con un mayor tamaño muestral.

Figura 6.16. Distribuciones estimaciones del precio del alquiler



Fuente: elaboración propia.

Para estudiar en detalle los valores numéricos asociados a la figura anterior, se han representado los parámetros principales del precio ponderado de alquiler en la Tabla 6.6. No se aprecian diferencias significativas, ni en los cortes por cuantiles ni en las medias, aunque los precios del modelo hedónico son ligeramente superiores a los originales. Por otra parte, tal y como se observaba en los precios de oferta, la mayor desviación típica del modelo final indica que se reduce, de forma general, la variabilidad final.

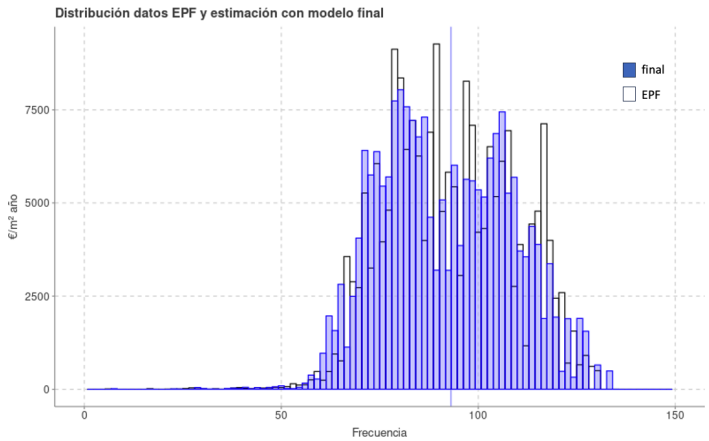
Tabla 6.6. Parámetros precios de alquiler anuales originales)

	EPF					Modelo Final				
Año	**Media**	**Desv**	**P25**	**P50**	**P75**	**Media**	**Desv**	**P25**	**P50**	**P75**
2011	99.61	571.13	81.15	100.77	117.81	99.74	507.76	83.39	102.27	118.74
2012	96.15	608.17	76.15	98.55	112.37	96.60	562.39	78.57	97.90	114.41
2013	91.27	463.69	74.49	93.33	108.16	91.54	433.12	76.14	93.61	108.87
2014	89.96	721.26	72.12	86.79	103.83	91.22	701.70	72.64	89.40	106.80
2015	85.78	388.83	71.91	86.55	100.24	86.11	336.69	72.80	88.38	101.09
2016	82.73	360.90	70.58	81.81	97.71	83.33	321.75	71.56	85.35	96.94
2017	86.18	421.04	73.64	85.35	99.53	86.54	373.09	73.20	88.34	100.89
2018	89.77	536.61	72.00	88.63	104.67	90.25	509.73	74.07	90.26	108.00
2019	94.65	546.02	75.82	95.79	109.48	95.63	535.06	78.21	95.63	111.92

Fuente: elaboración propia

De forma gráfica, la distribución de los valores de los distintos casos[18] se muestran en la Figura 6.17. Se observa que los valores estimados mantienen una distribución similar a los valores observados en la EPF, existiendo un muy ligero sesgo a infravalorar por parte del modelo. Existe una mayor irregularidad en la distribución, si se compara con la gráfica de oferta de la Figura 6.17, probablemente porque la fuente destino sea más irregular, similar a lo indica Kokot (2015) cuando compara la estabilidad de series de alquiler de precios de oferta y registros oficiales.

Figura 6.17. Distribución de precios de alquiler

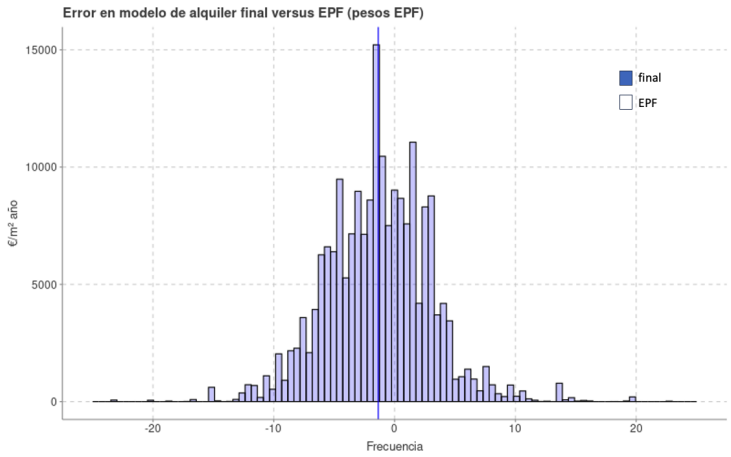


Fuente: elaboración propia.

[18]Para este caso densidad de precios ponderadas usando los pesos de la EPF, para estimar las funciones de distribución empíricas original y del modelo, en enfoque se basa una aproximación numérica inspirada en (Monahan, 2011).

Los errores del modelo se concentran en torno a cero, como se ve en la Figura 6.18, aunque al contrario del modelo de oferta, existe una menor concentración de errores nulos.

Figura 6.18. Distribución de errores del modelo de alquiler final



Fuente: elaboración propia.

Los valores de sesgo de los errores, mostrados en la Tabla 6.7, indican que el modelo final tiene un sesgo negativo (es decir que el modelo tiende a infravalorar). La desviación típica un 60% mayor que para el caso de la oferta, confirma la mayor dispersión de errores de la Figura 6.18.

Tabla 6.7. Sesgo modelos hedónicos de alquiler respecto a la EPF

Model	Media	Mediana	Pct. Media	Pct. Mediana	Desviación
Hedónico final	-1.53	-1.48	-1,542%	-1,494%	5.39

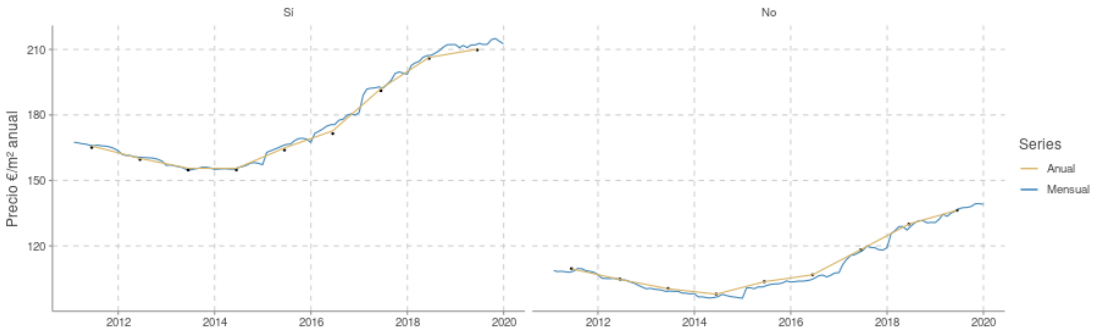
Fuente: elaboración propia

6.3.3 Coherencia temporal entre series anuales y mensuales

Puesto que las series mensuales finales de mercado se estiman con el modelo de mercado que se calcula sobre datos anuales, es necesario comprobar que esta diferencia de escalas temporales afecten a las series generadas. Cuando se concilian series con distintas frecuencias y múltiples periodos es habitual la presencia de discontinuidades (Hood, 2005), Chen y Andrews (2008) recogen que los cambios entre noviembre y febrero suelen ser particularmente superiores a los del resto del año .

Se puede comprobar gráficamente, en la Figura 6.19, que la variación entre los meses diciembre y enero es efectivamente superior a la observada en el resto del año.

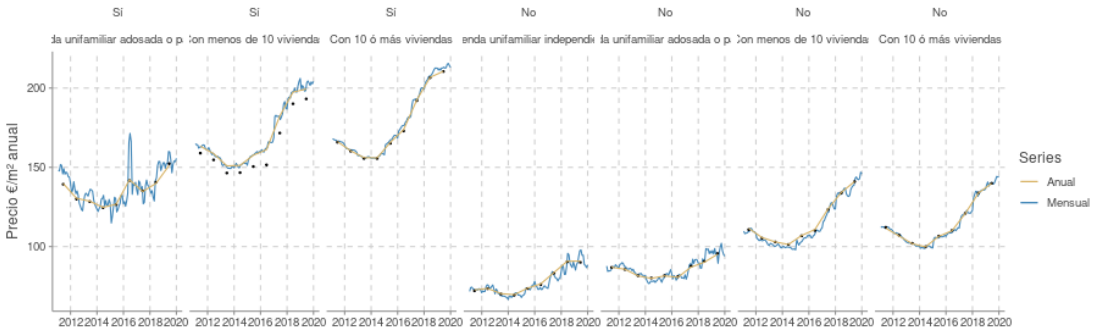
Figura 6.19. Precios de mercado precios mensuales y anuales, desglose capital y provincia



Fuente: elaboración propia.

El patrón anterior se acentúa a medida que se trabaja con estratos más pequeños, como muestra el desglose de precios medios por capital/provincia y tipo de edificio en la Figura 6.20. A este respecto, se puede establecer una relación con los resultados de los modelos de mercado, para los que cuanto menor es el tamaño del estrato más frecuente es la irregularidad de los resultados.

Figura 6.20. Precios de mercado precios mensuales y anuales, desglose tipo de edificio y capital/provincia



Fuente: elaboración propia.

Se puede confirmar el mismo comportamiento que con las series más agregadas, los cambios más abrupto en las series se corresponden, exactamente, a los periodos entre octubre y febrero, tal y como muestran los valores de la Tabla 6.8.

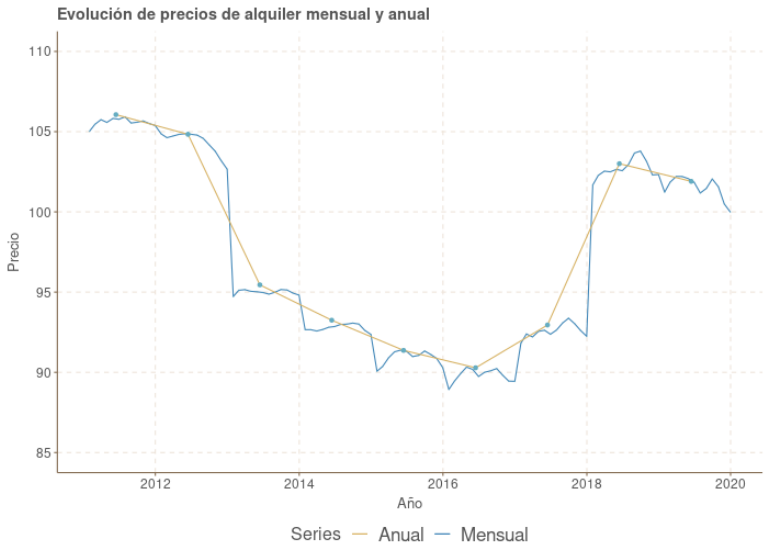
Tabla 6.8. Variación precio del alquiler con su mes anterior

Mes	Variación	Mes	Variación
Enero	3,32%	Julio	0,33%
Febrero	0,57%	Agosto	0,28%
Noviembre	0,52%	Septiembre	0,22%
Diciembre	0,42%	Abril	0,20%
Octubre	0,40%	Junio	0,20%
Marzo	0,39%	Mayo	0,18%

Fuente: elaboración propia

Estas discontinuidades no parecen atribuibles al proceso de reponderación, ya que, como se muestra la Figura 6.21, se observa de forma recurrente un cambio de escala de precios entre noviembre y enero del año posterior, para los pesos originales de la EPF y los calculados por la reponderación. Por otra parte las series mensuales y anuales son muy similares, con diferencias entre las medias mensuales y los valores anuales, que oscilan entre el 0,47% y el 1,48%.

Figura 6.21. Precios de mercado precios mensuales y anuales, viviendas plurifamiliares y pesos EPF



Fuente: elaboración propia.

Se puede concluir que los datos mensuales de mercado producidos por el modelo final reproducen las series anuales, aunque no exactamente, y tampoco capturan los cambios intermensuales, especialmente entre el fin y principio de año. Por tanto, como propone Eurostat (2015), es necesario un proceso de conciliación de series anuales y mensuales, para mitigar estos efectos indeseados.

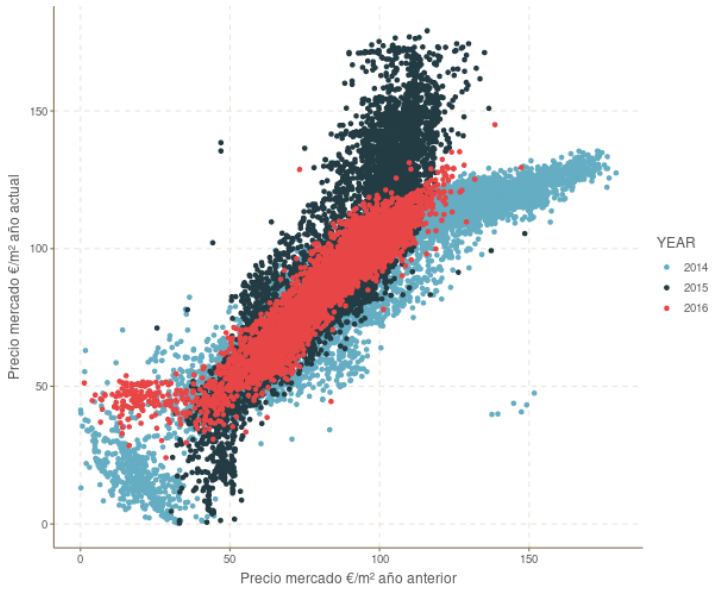
6.3.4 Capacidad de generalización para periodos futuros

Tal y como argumentan Shimizu *et al.* (2016), los modelos basados en datos de oferta cuentan con la ventaja de la inmediatez de la información. Sin embargo, aunque el dato del portal de internet está disponible de forma inmediata, el dato de mercado no, y se actualiza por el INE de forma anual y con un retraso de publicación de casi un año más. Esto resta, de forma estricta, aplicabilidad al modelo, al no permitir estimaciones de precio de mercado actualizadas.

Este problema se reduciría si la función que relaciona los precios del portal y los de alquiler fuera persistente en el tiempo[19], y por tanto, el modelo de un año sería válido para proyectar los datos de alquiler del año siguiente.

Para comprobar la hipótesis anterior, se estiman los precios de mercado para los años 2014, 2015 y 2016 usando el modelo de conversión correspondiente al periodo del año anterior. La Figura 6.22 demuestra gráficamente que existe una correlación entre ambos valores, en ella se relacionan el precio para un inmueble usando los modelos de años consecutivos. La pendiente variable según el año indica que relación varía ligeramente en el tiempo, pero sigue siendo fuerte.

Figura 6.22. Relación precios de mercado por el modelo de su periodo y el modelo del año anterior



Fuente: elaboración propia.

La muestra conformada por 27.766 observaciones ofrece un coeficiente de correlación de Pearson, entre años consecutivos, del 0,851, confirmándose la hipótesis de una fuerte relación persistente en el tiempo. Esta correlación es, en

[19] En este caso se mide la persistencia solamente entre años contiguos.

realidad, mucho más fuerte si observamos el desglose para cada año, como se aprecia en la Tabla 6.9, variando entre 0,86 y 0,94.

Tabla 6.9. Coeficiente de correlación de Pearson precio de oferta con modelos de años consecutivos

Año	Coef. correlación de Pearson
2014	0.91
2015	0.86
2016	0.94

Fuente: elaboración propia

Para confirmar que la relación entre magnitudes es significativa, se desarrolla un modelo lineal de mínimos cuadrados ordinarios que estima el precio de alquiler del año en curso, según el precio estimado del año anterior, y definido como:

$$\hat{P}_e(t) = \beta_0 + \beta_1 P_e(t-1) + \sum_{a=1}^{N} \gamma_a D_a + \varepsilon_t \qquad [6.10]$$

donde $\hat{P}_e(t)$ representa el precio de alquiler estimado para un estrato e y un año t, D_a es una variable *dummy* que toma valor 1 si el año t se corresponde a a, y cuya función es controlar las variaciones específicas de cada año. Finalmente, ε_t se refiere al término de error aleatorio.

La Tabla 6.9 muestra los coeficientes del modelo, donde *ANNUAL_PR_RENT_T1* se corresponde a β_1, mientras que *YEAR2015* y *YEAR2016* se refiere a los coeficientes γ_a de las *dummy* de tiempo D_a en la expresión [6.10]. Se observa que los *p-valores* de todos los coeficientes de las variables independientes (Tabla 6.10) son significativos, con un grado de significación inferior al 0,001. Lo cual confirma que las variables independientes estudiadas aportan información relevante y valiosa para explicar y predecir el comportamiento de la variable dependiente.

Tabla 6.10. Modelo de relación precios estimado con modelo del año anterior

	Estimación	**std.error**	**t value**	**Pr(>|t|)**	**signif.**
(Intercept)	9.50	0.31	31	< 2e-16	***
ANNUAL_PR_RENT_T1	0.88	0.00	286	< 2e-16	***
YEAR2015	9.86	0.19	52	< 2e-16	***
YEAR2016	4.99	0.22	23	< 2e-16	***

Fuente: elaboración propia

Códigos signif.: *** 0,001 ** 0,01 * 0,05 . 0,1 1
Error estándar de los residuos: 14 sobre 27766 grados de libertad (DF)
R^2: 0,74893, R^2 ajustado: 0,74891
Estadístico-F: 27605 sobre 3 y 27762 DF, p-value: < 2,2e-16
Num. observaciones: 27766

En este capítulo, se ha presentado un método que corrige eficazmente el sesgo producido por la falta de información zonal del modelo de mercado. Los precios ajustados se utilizarán para el cálculo de las series de precios de los distintos estratos de la población, que posteriormente se aplicarán en la construcción del índice de precios final, presentado en el Capítulo 8. Sin embargo, los modelos calculados hasta este momento ofrecen datos anuales, por lo que se desarrollará un método para desagregar los datos a frecuencia mensual, y que se presentará en el Capítulo 7.

Anexo 6a. Identificación y corrección de sesgos

Las estimaciones de un modelo de aprendizaje automático de árboles tiende a ser insesgado en el sentido de que la suma de los errores (observados contra los estimados) es cercano a cero (Hastie *et al.*, 2017). Sin embargo, los modelos de regresión calculados con estos métodos pueden arrojar resultados sesgados en un sentido diferente (Zhang y Lu, 2012): los valores pequeños se sobreestiman y los valores altos se infraestiman. Para muchos propósitos es importante cualificar correctamente los casos extremos de la distribución. En el caso de las valoraciones inmobiliarias existe una gran diversidad de mercados, y dentro de un submercado existen inmuebles singulares, por tanto este tipo de modelos deben ser capaces de tratar los casos más extremos.

En un modelo de regresión el error del modelo es la suma de varianza aleatoria (ruido blanco o ε) del sesgo del predictor y la varianza del predictor, donde los dos últimos componentes se denominan riesgo de la función de regresión, descrito en [6.11]. Breiman (1996) demuestra que la técnica de *bagging* puede reducir de forma eficaz la varianza del predictor, pero no actúa sobre el sesgo, siendo este último el factor dominante del riesgo[20] de los modelos. Por otra parte, la reducción de la varianza da lugar a que este tipo de métodos no traten correctamente los valores extremos y por tanto pueda ser necesaria una corrección de sesgo.

$$R[\hat{f}(x)] = Bias[\hat{f}(x)] + \sigma^2[\hat{f}(x)] \qquad [6.11]$$

donde $R[\hat{f}(x)]$ se refiere al riesgo del estimador $\hat{f}(x)$, $Bias[\hat{f}(x)]$ es su sesgo, y $\sigma^2[\hat{f}(x)]$ la varianza del mismo.

Existe una segunda fuente de sesgo de los modelos estadísticos producida por las transformaciones de la variable dependiente. En nuestro caso, se transforma la variable de precio a escala logarítmica, lo que es útil en los modelos lineales para reducir la heterocedasticidad en la variable respuesta, pero puede dar lugar a sesgo en la magnitud estimada al revertir la transformación de la variable dependiente.

Se puede establecer una medida numérica de sesgo como la diferencia entre los valores observados, $\hat{f}(observado)$, y los valores estimados por el modelo, $\hat{f}(estimado)$. Por tanto, se define el error de un estimador $\epsilon[\hat{f}(x)]$ como:

[20] El riesgo mide los dos aspectos clave en el ajuste de los modelos: la varianza y el sesgo.

$$\epsilon[\hat{f}(x)] = \hat{f}(observado) - \hat{f}(estimado) \quad [6.12]$$

Y el sesgo como esperanza de los errores del modelo en:

$$Bias[\hat{f}(x)] = E\{\epsilon[\hat{f}(x)]\} \quad [6.13]$$

En general, el control de sesgo se afronta desde dos perspectivas, el primero basado en la escala del punto (*point-scale*) o basado en la escala de la distribución (*distribution-scale*). El primer enfoque ajusta los valores estimados para que la desviación entre la estimación y el valor observado sea mínimo, y el segundo ajusta la distribución de la variable estimada de forma que las distribuciones acumuladas, del modelo y la variable dependiente, se correspondan. Las correcciones de sesgo sobre los valores puntuales se han tratado de forma extensa en la literatura estadística (Giffen *et al.*, 2022; Hort *et al.*, 2022; Pagano *et al.*, 2022)

Para *Random Forests* el modo más sencillo de controlar el sesgo es a través los registros OOB, descritos en detalle en el Anexo II del Capítulo 3. Por ejemplo, Liaw y Wiener (2002) proponen usar una regresión lineal sobre este conjunto datos para realizar el ajuste, o Zhang y Lu (2012) propone usar una regresión generalizada con la muestra completa. Existen otros método alternativos que usan un segundo modelo de *Random Forests* para controlar el efecto del sesgo, denominados *"one-step boosted forests"* (OSFB) (Zhang y Lu, 2012), que en general muestran mejores resultados que los método basados en los registros OOB.

Los método basados en la corrección de la distribución parten de encontrar un modelo de correspondencia entre las dos distribuciones, y no se han aplicado tan extensamente en todos los campos, siendo quizá la meteorología el único campo donde se ha aplicado de forma más frecuente.