revistainvi

Volumen 39, nro. 110, mayo 2024

ISSN 0718-1299

Artículos

Metodología de estimación del tiempo de venta a través de modelos de supervivencia

Recibido: 2023-02-13

Aceptado: 2023-11-20

David Sánchez-Cabrera

Universidad Católica Santa Teresa de Jesús de Ávila, Ávila, España,
david.sanchez@ucavila.es
https://orcid.org/0009-0009-5034-9300

Julio González-Arias

Universidad Nacional de Educación a Distancia, Madrid, España,
jglez@cee.uned.es
https://orcid.org/0000-0003-4993-4739

David Rey-Blanco

Idealista, Madrid, España, drey@idealista.com
https://orcid.org/0000-0002-3549-1714

Cómo citar este artículo:

Sánchez-Cabrera, D., González-Arias, J., y Rey-Blanco, D. (2024). Metodología de estimación del tiempo de venta a través de modelos de supervivencia. *Revista INVI, 39*(110), 164-202.
https://doi.org/10.5354/0718-8358.2024.69772

Metodología de estimación del tiempo de venta a través de modelos de supervivencia

Resumen

El conocimiento detallado de aquellos factores que presentan mayor relevancia en la probabilidad de venta de las viviendas puede resultar de suma importancia al tratarse de una inversión a largo plazo. Tras identificar una serie de bloques que marcan la evolución de la venta de las mismas (las características internas del inmueble en cuestión, su localización, su grado de sobreprecio y, sobre todo, un interés real por parte del comprador), se presenta un estudio novedoso que modeliza la probabilidad de venta de los inmuebles a lo largo del tiempo mediante técnicas de *machine learning* aplicadas a problemas de supervivencia, logrando valores de *C-index* de 76% y 72% en chalés y pisos, respectivamente. El proceso metodológico se ha testado sobre la capital de España, Madrid, a partir de los datos recopilados desde la principal plataforma de mercado del país durante el periodo 2018-2019. Estos datos fueron ponderados según información oficial, pero la metodología es escalable a cualquier municipio. No solo vendedores, compradores o intermediarios pueden verse beneficiados con este aporte, sino también los agentes públicos, de cara a tomar decisiones enfocadas al diseño o prevención en el tema de la vivienda.

Palabras clave: análisis de supervivencia aplicado a la venta de viviendas, días en mercado (DEM), mercado de la vivienda, precio y valoración de la vivienda, probabilidad de venta en función del tiempo.

Sales Time Estimation Methodology Through Survival Models

Abstract

Knowing in detail those factors that are most relevant to the probability of selling homes can be of utmost importance, among other things, as it is a long-term investment. After identifying a series of blocks that mark the evolution of their sale, such as the internal characteristics of the property in question, its location, its degree of overpricing and, above all, a real interest on the part of the buyer, a novel study is presented that models the probability of selling properties over time using machine learning techniques applied to survival problems, achieving C-index values of 76% and 72% in chalets and apartments, respectively. The methodological process has been tested in the capital of Spain, Madrid, based on data collected from the country's main market platform during the 2018-2019 period, weighted according to official information, but the methodology is scalable to any municipality. Not only sellers, buyers or intermediaries can benefit from this contribution, but also public agents, in order to make decisions focused on design or prevention in the area of housing.

Keywords: housing market, price and valuation of the home, survival analysis applied to home sales, time-dependent selling probability, time on market (TOM).

Introducción

En toda transacción de compraventa de un inmueble, el objetivo de cada agente es diferente, pues el vendedor ha de encontrar un equilibrio entre el tiempo que se demora la venta de un inmueble y el precio recibido por este (Anglin *et al.*, 2001); por otro lado, el comprador espera conseguirlo al mínimo precio posible, pudiendo retrasar incluso la compra anticipando mayores caídas en los precios, hasta el punto de convertirse en un "comprador depredador" (Selcuk, 2012). Cualquier tercero, ya sean agencias físicas o plataformas de mercado, ayudarán a que la transacción se llegue a consumar. Ayudar a todos ellos a conocer patrones de comportamiento les facilitará la toma de decisiones.

Por ello, encontrar una armonía entre las características/precio de un bien y el tiempo que estará disponible en el mercado (*time on market*, en adelante TOM) es primordial. Diversos autores, como se verá en el apartado siguiente, han encontrado relaciones entre el TOM y otras variables, ya sean características endógenas o exógenas a las propias viviendas, que han sentado las bases de este artículo y han permitido ir un paso más allá. Por ejemplo: en propiedades con cualidades superiores y a un precio inferior al promedio de mercado, el tiempo de exposición es considerablemente menor que el de una propiedad en condiciones de mercado estándar.

El presente trabajo, elaborado a partir de datos publicados en régimen de venta de una plataforma de mercado –en concreto, la plataforma Idealista, portal web inmobiliario más visitado en internet en España, con más de 40 millones de visitas al mes según se publica en su web idealista.com–, presenta una metodología para desarrollar un modelo centrado en el estudio del balance entre las características de los inmuebles y el tiempo que tardan en venderse. Para desarrollar el análisis, se emplean datos de ofertas en el portal web de España, dado que es uno de los países europeos con mayor propensión a la compra de vivienda por parte de su población, en torno al 77% en 2017 (Instituto Nacional de Estadística [INE], 2019), muy por encima de países como Reino Unido (65%), Francia (64%) o Alemania (51%) (Eurostat, 2019). Por ello, es recomendable que en una inversión de tal magnitud –en términos de plazo e importe– el comprador conozca en detalle aquellos factores más relevantes en la probabilidad de venta de los inmuebles.

Para trabajar con una mayor y más representativa muestra, este estudio se ha centrado en el Madrid de los años 2018 y 2019, evitando así tomar los registros de 2020 y 2021, dado el anómalo comportamiento registrado como consecuencia del COVID-19. Se ha considerado como fecha de puesta en mercado la fecha de alta en el portal, y se asume que la fecha de baja se refiere a la fecha de venta del inmueble, estableciendo un año (como máximo) como periodo de estudio para evitar sesgos derivados de comportamientos estacionales[1].

1 Ngai y Tenreyro (2014) demostraron que el comportamiento de la venta de viviendas presenta diferencias según la época del año, hecho observado también en España (INE, 2023a).

Es fundamental conocer el comportamiento de todos los agentes involucrados en una transacción de compraventa, especialmente en modelos que buscan armonía entre las características técnicas y el tiempo. Por una parte, se encuentran los vendedores y compradores: mientras los primeros desean conocer información que les permita optimizar la venta (precio de publicación, necesidad de reforma de la vivienda y estilo de publicación, entre otros elementos), los segundos podrán jugar con la probabilidad de venta del bien interesado para elegir la "mejor vivienda entre las posibles", así como el momento idóneo para lograrla (Bhuiyan y Hasan, 2016). Por otra, entran en escena las agencias o portales inmobiliarios, al tratarse de intermediarios en el proceso: a través del modelo aquí presentado podrían asesorar al vendedor, ofreciéndoles un servicio adicional en su software, para optimizar el proceso de venta (como puede ser ajustar el precio o mejorar las características de la comercialización).

El objetivo principal del trabajo es encontrar un modelo final que, a través de un estudio exploratorio con enfoque cuantitativo, determine la probabilidad de venta de las viviendas en función del tiempo. Trata de ser una innovación, tanto en los datos (volumen, características y soporte de partida de ellos) como en su aplicación y se estimará mediante el uso de técnicas de *machine learning*, determinando los atributos que tienen más importancia. Este objetivo central traerá consigo unos objetivos secundarios, como: la aplicación de procesos en la reponderación de los datos para dotar de consistencia a la muestra según fuentes oficiales –cuestión que tiene crucial importancia– y que permite soportar las fuentes de datos privadas, en las que pudiera haber distorsiones, sobre la estadística oficial; la estimación del precio teórico de los inmuebles para poder calcular el sobreprecio de cada uno de ellos; y la generación de unas puntuaciones que definan de un modo notorio cada uno de los barrios.

Problemática y estado del arte

En el mercado de la vivienda, para todo aquel interesado en él, resulta esencial encontrar un equilibrio entre el precio de venta y el TOM, teniendo en cuenta que un elevado importe de publicación para un determinado inmueble incrementará su tiempo de venta: al necesitar una mayor revisión del valor inicial, conllevará un mayor tiempo alcanzar su valor idóneo de mercado (Anglin *et al.*, 2001), generando un acuerdo que se cerrará por un monto inferior en relación a aquél con una estimación correcta en su inicio (Knight, 2002). Pero, además, si la propiedad es atípica (por tener, por ejemplo, un precio excesivamente alto dentro de los portales) permanecerá todavía mayor plazo en el mercado (Haurin, 1988). De todos modos, para paliar este efecto y conseguir así una mayor rapidez en la venta, un vendedor puede acortar el plazo en la medida que aumente la rebaja en la potencial revisión (Khezr, 2015).

Diversos autores han estudiado relaciones entre el tiempo que tarda un bien en venderse y otras variables, observando ciertos patrones. En primer lugar, en sendos estudios llevados a cabo en los años noventa sobre viviendas unifamiliares en Texas (Estados Unidos), se observó que el plazo de venta se incrementaba cuanto menor era la antigüedad del inmueble o mayor su tamaño (Forgey *et al.*, 1996), así como con un mayor número de plantas del inmueble, mientras que la presencia de piscina o garaje lo reducía (Anglin *et al.*, 2001). También en la década de los noventa y en Estados Unidos, concretamente en Boston, se halló una significativa relación entre el valor hipotecado y la tasación (*loan-to-value*, en adelante LTV), de modo que los inmuebles con un alto LTV presentan mayor dificultad para venderse (Genesove y Mayer, 1997).

Por otro lado, en una investigación llevada a cabo sobre edificios residenciales de gran altura en Singapur en los primeros años del 2000, se percibió que aquellos inmuebles con una orientación este/oeste o plantas más bajas tenían un mayor tiempo de venta (Li, 2015). En este sentido, se afirmaba que la localización juega un factor importante según se aprecia en los análisis llevados a cabo en la primera década de los 2000 sobre la capital de Eslovenia (Cirman *et al.*, 2015) o la costa este de Estados Unidos (Johnson *et al.*, 2007); por ejemplo, la proximidad a centros educativos de calidad estaba ligada con la probabilidad de venta, según un estudio similar al propuesto sobre diversos inmuebles de Indiana (Estados Unidos), obtenidos en 2015 del portal inmobiliario estadounidense Trulia (Bhuiyan y Hasan, 2016).

Otras consideraciones relevantes incluyen tanto las motivaciones del vendedor como el modo de actuación en épocas con diversas coyunturas económicas. En cuanto al primer punto, si existe una urgencia real por vender, habrá disposición a vender más barato (Quan y Quigley, 1991) ya que "los vendedores más motivados pueden reducir el tiempo de venta hasta en un 30%" (Glower *et al.,* 1998), siendo el coste de mantener la propiedad un aspecto para tener en cuenta (Geltner *et al.,* 1991). En lo que se refiere a la influencia del mercado, tanto la inflación como el precio de los bonos están directamente relacionados con la probabilidad de venta (Scofield y Devaney, 2017); además, si los tipos de interés son altos, interesa esperar para obtener precios más elevados, mientras que una venta rápida en tiempos de tipos de interés bajos es más aconsejable

(Kang y Gardner, 1989). Esta teoría es apoyada por An *et al.* (2013), quienes comentan que la relación entre el TOM y el precio de venta depende de las condiciones del mercado. Esta relación generalmente es directa (McCall y Lippman, 1984), pero se convierte en inversa en época de crisis (Lazear, 1986).

Por otra parte, es relevante comentar que el precio anunciado en las publicaciones se ha convertido en una variable fundamental, pero en el mercado de la vivienda resulta imprescindible conocer la diferencia entre el precio de anuncio y el de venta: mientras el primero es simplemente el reflejado producto de anunciar que un determinado inmueble está disponible para su comercialización (pudiendo ser no único a lo largo del tiempo en el caso de sufrir variaciones), el segundo es aquel con el que se cierra finalmente la transacción (no teniendo por qué coincidir con el valor del inmueble, pues este es "el mejor precio razonablemente obtenible por el vendedor y el más ventajoso razonablemente obtenible por el comprador" (International Valuation Standards Council, 2022), el cual es dependiente de condiciones cíclicas, como la tasa de interés, coste de los materiales –en el caso de las viviendas usadas, para una posible reforma a llevar a cabo– y valor de las acciones de las empresas de construcción, como se aprecia en un estudio llevado a cabo en Santiago de Chile entre los años 2009 y 2019 (Vergara-Perucich, 2021). Sin embargo, Han y Strange (2016) muestran que no se trata de algo inexorable a cumplir específicamente en el proceso de compra ni siquiera de un techo en el precio de venta –como sí afirman entre otros autores Chen y Rosenthal (1996)–, sino más bien un parámetro sobre el que se relaciona el precio de venta –el ratio entre ambos suele variar a lo largo del tiempo, situándose el precio de venta por debajo en la década de los noventa y por encima a partir del año 2000 (Ahmed *et al.*, 2016)– y, sobre todo, un filtro inicial para seleccionar posibles compradores. Este valor marcará en mayor o menor medida el posible efecto del sobreprecio del inmueble (*degree of overpricing,* DOP, calculado como la relación entre el precio publicado y el teórico del inmueble). Anglin *et al.* (2001) manifestaban que estas viviendas tendrían un plazo de venta mayor, aunque Bich *et al.* (2020), observaron una disminución, responsabilizando a la escasa oferta, del mismo durante las primeras semanas en su estudio entre 2017 y 2018 en Ho Chi Minh (Vietnam).

A continuación, se sintetizan, según su fecha de publicación, las diferentes corrientes comentadas con relación al tiempo de venta (en el cuadro resumen de la Tabla 1), así como los resultados alcanzados para los estudios que disponen de métricas que miden la bondad de los ajustes (Tabla 2).

Tabla 1.
Sinopsis de argumentos que relacionan variables con un mayor tiempo de venta de los inmuebles.

Autores	Variable	TOM
McCall y Lippman (1984) / Lazear (1986)	Mayor precio de venta	Mayor / menor en épocas de crisis
Haurin (1988)	Propiedad atípica	Mayor
Forgey *et al.* (1996)	Propiedades nuevas o de mayor tamaño	Mayor
Genesove y Mayer (1997)	Alto LTV	Mayor
Anglin *et al.* (2001)	Mayor precio de publicación inicial	Mayor
Anglin *et al.* (2001)	Presencia de piscina o garaje	Menor
Anglin *et al.* (2001) / Bich *et al.* (2020)	DOP	Mayor / menor
Khezr (2015)	Mayor rebaja de precio en las revisiones de éste	Menor
Li (2015)	Orientación este/oeste o plantas bajas	Mayor
Cirman *et al.* (2015)	Localización central (diferencia por zonas)	Mayor
Bhuiyan y Hasan (2016)	Proximidad a la calidad de los centros educativos	Menor
Scofield y Devaney (2017)	Mayor inflación y precio de los bonos	Mayor

Fuente: elaboración propia.

Tabla 2.
Mejores métricas obtenidas por cada uno de los autores que estimaron el TOM.

Autores	Número de variables empleadas	eMétrica evaluada y mejor valor
Forgey *et al.* (1996)	14	R^2 = 30%
Genesove y Mayer (1997)	12	*Log Likelihood* = –3.542,1
Anglin *et al.* (2001)	20	*Log Likelihood* = –2.813,69
Khezr (2015)	3	R^2 = *26%*
Li (2015)	25	*Log Likelihood* = –1.966
Cirman *et al.* (2015)	8	*Log Likelihood* = –315,20
Bhuiyan y Hasan (2016)	8	*C – index = 69%*
Scofield y Devaney (2017)	21	*Log Likelihood* = –9.311
Bich *et al.* (2020)	15	*Wald* $_X2$ = 90,67%

Fuente: elaboración propia.

En los trabajos mencionados se han aplicado técnicas estadísticas y matemáticas clásicas (regresiones o de análisis de supervivencia). Este estudio propone una metodología de análisis que aproveche el potencial que ofrecen los algoritmos derivados del *machine learning*, los cuales no requieren, en su mayoría, análisis de autocorrelación, pruebas de normalidad o heterocedasticidad, entre otras, por ser métodos no paramétricos. Este proceso en la búsqueda de modelos más robustos ya ha empezado a desarrollarse en investigaciones previas como las de Choy y Ho (2023), los cuales batieron claramente a los modelos tradicionales en la estimación de precios por más de 10pb, o el elaborado por Baldominos *et al.* (2018), con objeto de la detección de oportunidades de viviendas en tiempo real, por cotizar por debajo de su precio esperado.

Finalmente, conviene señalar que la mayor parte de los trabajos que se han reseñado parten del estudio directo de muestras compuestas por datos de una dimensión reducida, en varios casos ofrecidos por plataformas, sin considerar el sesgo que podría derivarse. Así mismo, tampoco se confrontaron estos datos con los datos oficiales de compra/venta, cuestión que se convierte en crítica, más en un sector como este y dado el origen virtual de los mismos, que apunta reducir los posibles errores que los modelos esconderían. Esta práctica, habitual en estudios de la Unión Europea, conforme a los criterios metodológicos establecidos por Eurostat –como por ejemplo en la Encuesta de Población Activa desde la actualización de 2002 (INE, 2002a)– será tratada en este estudio.

Sobre esta base, el presente análisis tiene como objetivo central estimar la probabilidad de venta de las viviendas en función del tiempo, considerando las derivadas de los modelos TOM, con otras cuestiones que, como ya se ha apuntado en estudios previos, pueden tener incidencia en el hecho de la consumación de la operación comercial.

Metodología

Previamente al desarrollo metodológico, se tomaron una serie de decisiones en el proceso de pre-procesado inicial:

- No todos los anuncios que se dejan de publicar son considerados como vendidos, sino solo los que llevan un indicador de desactivación específico (básicamente los no fraudulentos). Se considera como fecha de inicio la primera aparición del inmueble en el portal, mientras que por la última se supondría la venta.
- El periodo de estudio se establece en un año, evitando así la estacionalidad –la cual viene derivada de la concentración de las transacciones en ciertos momentos del año– por lo que se considera que la mejor forma de abordar series más amplias es en este espacio de tiempo. Así mismo, se eliminan aquellos inmuebles que se suponen vendidos en su primera semana de vida (hecho que es muy poco probable y no reflejará la realidad de la población).

El proceso metodológico propuesto se desarrolla en dos fases: en una primera se da consistencia y veracidad a la muestra de origen, por medio de procesos de re-muestreo, para, en una segunda fase posterior, aplicar técnicas avanzadas en estadística y matemática que permitan desentrañar las relaciones significativas de las variables y concluir sobre el objetivo marcado:

- *Fase de re-muestreo:* los datos de origen son susceptibles de tener importantes sesgos. Desde el comienzo, son creados por los propios interesados y subidos a una plataforma, lo que en el proceso puede generar imprecisiones. Pero, adicionalmente, en ocasiones se aplican estrategias comerciales que pueden suponer la duplicidad de registros con pequeñas modificaciones. Del mismo modo, no siempre se puede constatar o asumir que la desaparición de un anuncio suponga la realización final de una compra/venta. En este sentido, se requiere de un proceso de re-muestreo sobre datos oficiales, que permita soportar en origen la muestra inicial de partida, y dar veracidad al resto de procesos de análisis que posteriormente se desarrollarán. Por tanto, es necesario adaptar los registros que se disponen a la estadística pública, de modo que se ponderen estos en función de los datos oficiales, es decir, los inmuebles de la muestra tendrán unos pesos para adecuarse totalmente a las transacciones realmente efectuadas.

- *Fase de análisis:* establecer la variable tiempo como objeto del análisis –y correlacionarla con un conjunto de características de los inmuebles– determinará los modelos y algoritmos de supervivencia aplicados, identificando el que muestre un mejor comportamiento. Para ello, previamente se realiza una estimación del precio teórico de los inmuebles, así como la generación de unas puntuaciones que simbolicen los diferentes barrios disponibles.

Comenzando por la primera fase, es de destacar que las estimaciones llevadas a cabo en la realización de un proceso de selección de datos por muestreo, como el llevado en el presente trabajo, pueden sufrir ciertos sesgos, ya sea por falta de respuesta (afectando de manera desigual a los distintos grupos de la población) o por sobre-representar a determinados colectivos (por ejemplo, por un desfase temporal) (Deville *et al.*, 1993). Para minimizarlos, una opción usual consiste en utilizar fuentes estadísticas externas fiables y re-ponderar la muestra a través de los factores de elevación, siendo la re-ponderación aquel proceso que trata de corregir los pesos o factores de elevación originales para lograr unos factores finales tales que, al aplicarlos, la estimación de las variables para las que se dispone de información de una fuente externa fiable (datos de referencia para la re-ponderación) coincidan con los datos de dicha fuente.

El procedimiento de la re-ponderación necesita partir de unas variables explicativas (en este caso: ser vivienda usada, superficie de la misma y barrio en el que se localiza) que existan tanto en la muestra como en una fuente estadística oficial que la soporta –ya sea un Censo, el Padrón o un Registro Administrativo– y que presenten una correlación lo más fuerte posible con las variables de interés (aquellas cuyas estimaciones revisten la mayor importancia de la muestra, siendo en el estudio la venta o no venta del inmueble), conforme al INE (2002b).

Si se parte de una muestra de tamaño n, llamando w_{nx1} al vector fila de pesos originales y w' al vector homólogo de pesos transformados, cualquier procedimiento de re-ponderación que se aplique dará lugar a una relación del tipo siguiente (es decir, los nuevos pesos serán tanto en función de los originales como de las variables auxiliares elegidas):

$$'w' = w'(x, X) \qquad (1)$$

La información auxiliar proporcionada por la muestra va a estar representada en la matriz Xnxp' donde cada fila contiene los valores de las variables auxiliares para cada uno de los individuos de la muestra. Los nuevos pesos habrán de cumplir la condición de equilibrado de la muestra:

$$X'w' = x \qquad (2)$$

Siendo x el vector de efectivos poblacionales proporcionados por las fuentes externas utilizadas. Con los pesos w', se pasaría a obtener las nuevas estimaciones para cualquier variable de interés Y comentada.

Todo este proceso de re-muestreo es necesario para obtener los pesos que deberían tener los inmuebles de la muestra, los cuales, pese a ser diferentes en inicio, deben ser iguales a los de la estadística pública, y que no indican otra cosa que cuanto representa cada uno sobre la población real.

En nuestro caso particular, y persiguiendo lo comentado en el párrafo anterior para dotar de consistencia a la muestra de Idealista, esta será re-ponderada en función de los datos oficiales de compraventa publicados por el Ayuntamiento de Madrid (2020) –estadística regional que desglosa las viviendas entre nuevas y usadas por cada uno de los barrios, así como en función del tamaño de ellas–. La necesidad de este proceso queda reflejado en la comparación de los datos obtenidos: existían discrepancias tanto a nivel general (por ejemplo, en 2019 hubo 90.887 ventas teóricas según el portal frente a las 34.066 oficiales, lo cual permite pensar que se retiran del portal más de dos de cada tres inmuebles sin haber llegado a concretar una venta real) como por barrios (en 2018, para "Atocha" se disponían de 12, todas entre 60 y 80 m^2 vs. dos en el registro oficial, una inferior a 60 m^2 y la otra superior a 100 m^2). Estos aspectos quedan recogidos en la Tabla 3.

La validad, veracidad y credibilidad del modelo final es, por tanto, soportada por una base estadística oficial a través de un proceso de calibración de la muestra obtenida a los datos oficiales para cada año. De esta forma, se obtienen pesos específicos con los que se re-ponderará la muestra del portal conforme la razón que guardan con la realidad observada, asumiendo para los inmuebles no vendidos el peso obtenido para los vendidos, manteniendo así proporcionalidad. Este proceso de muestreo es crucial para mantener la estabilidad temporal y espacial.

Ahondando ahora en los aspectos técnicos que darán soporte al trabajo, la base del mismo es el análisis de supervivencia, el cual incorpora un conjunto de técnicas estadísticas que proporcionan una estimación del tiempo hasta que ocurre un determinado evento, generalmente denominado como "muerte" (Kleinbaum y Klein, 1996). Sea $T \geq 0$ la variable aleatoria que representa el tiempo de supervivencia de un sujeto y $\delta = \{0,1\}$ una variable aleatoria que indica la "muerte" o no (censura) de un determinado suceso:

$$\delta = \begin{matrix} 1, \textit{ si hay muerte} \\ 0, \textit{ en otro caso} \end{matrix} \qquad (3)$$

Surgen entonces dos conceptos claves. En primer lugar, la función de supervivencia, la cual representa la probabilidad de que un determinado individuo sobreviva más allá de un tiempo t:

$$S(t) = P(T > t) \qquad (4)$$

Y la función de riesgo (*hazard function* en inglés), la cual, suponiendo supervivencia hasta el momento t, mide el cambio potencial de muerte en t por unidad de tiempo:

$$h(t) = \frac{P(t \leq T < t + \triangle t | T \geq t)}{+\triangle t} \qquad (5)$$

Siendo la función de riesgo acumulada:

$$H(t) = \int_0^t h(x)dx \qquad (6)$$

Donde la relación entre la función de supervivencia y esta última viene dada por la ecuación:

$$S(t) = e^{-H(t)} \qquad (7)$$

Los atributos definitivos, denotados por un vector , serán combinados a través de una función para generar un modelo final que proporcionará la supervivencia en un determinado momento . Este modelo se puede representar a través de una ecuación general del tipo (donde marcará la función de supervivencia base):

$$S(t,x) = S_0(t)e^{f(x)} \qquad (8)$$

Determinar la probabilidad de venta de un inmueble a lo largo del tiempo en base a sus características no es más que la aplicación de esta teoría al caso de uso concreto del mercado de la vivienda. Por esta razón, se estimaron diferentes modelos relativos al análisis de supervivencia aplicados a la venta de los inmuebles, para así poder construir curvas de probabilidad de venta de los inmuebles, entrenando los modelos con el 80% de la muestra total a través de *cross-validation* (CV) y reservando el 20% restante para el test: los métodos tradicionales no paramétricos como Kaplan-Meier, el cual se basa en el supuesto de que los grupos son homogéneos con respecto a sus covariables (Kaplan y Meier, 1958); los modelos paramétricos, que asumen distribuciones de probabilidad conocidas para estimar el tiempo hasta la ocurrencia del evento de estudio, como el de Weibull, log-logistic y log-normal (Oxford Spring School, 2007); los semiparamétricos, como la regresión de Cox, denominados así porque contienen una parte no paramétrica –denominada riesgo basal– y otra paramétrica –dada por una exponencial para representar un conjunto de covariables–, los cuales asumen proporcionalidad de los riesgos de los individuos a lo largo del tiempo (Cox, 1972); y, finalmente, aquellos provenientes del *machine learning*, como es el Árbol de Decisión o un conjunto de ellos, el *Random Forest* (Pölsterl, 2020). Se consideran más apropiadas estas técnicas que otras llevadas a cabo en varios de los estudios comentados del apartado anterior, como los modelos *logit* o *probit* para predecir el tiempo en el mercado, principalmente porque pierden información de los registros censurados (aquellos que no han presentado un evento positivo durante el periodo de estudio).

Se muestra un cuadro resumen en la Tabla 4 de los algoritmos utilizados para la predicción del tiempo de venta, ordenadas por fecha de publicación, a modo de comparación.

Tabla 3.

Datos de inmuebles oficiales (O) y el portal (P), así como la ponderación a aplicar (W), por barrios, superficie y año.

Barrio	2018												2019											
	< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*			< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*		
	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W
San Cristóbal	127	37	3,43	110	176	0,63	7	12	0,58	4	10	0,40	147	36	4,08	124	225	0,55	16	20	0,80	6	11	0,55
Butarque	98	24	4,08	52	51	1,02	38	80	0,48	29	65	0,45	105	27	3,89	65	68	0,96	78	95	0,82	288	88	3,27
Ángeles	84	26	3,23	115	130	0,88	133	133	1,00	53	88	0,60	72	37	1,95	227	229	0,99	95	152	0,63	299	124	2,41
Los Rosales	136	55	2,47	91	133	0,68	44	65	0,68	27	78	0,35	146	92	1,59	86	170	0,51	41	117	0,35	18	125	0,14
Villaverde Alto - Casco Histórico de Villaverde	262	161	1,63	177	635	0,28	86	278	0,31	112	347	0,32	194	205	0,95	172	625	0,28	36	315	0,11	36	311	0,12
Orcasitas	2	3	0,67	22	14	1,57	71	64	1,11	96	113	0,85	6	11	0,55	10	15	0,67	53	75	0,71	63	223	0,28
Ensanche de Vallecas	334	192	1,74	249	173	1,44	265	281	0,94	196	216	0,91	268	206	1,30	193	171	1,13	265	281	0,94	359	235	1,53
Buenavista	202	120	1,68	134	251	0,53	56	132	0,42	36	117	0,31	152	151	1,01	81	324	0,25	57	198	0,29	70	177	0,40
Cuatro Vientos	7	0	-	16	4	4,00	6	1	6,00	2	2	1,00	14	1	14,00	7	1	7,00	3	3	1,00	4	1	4,00
San Fermín	50	9	5,56	70	67	1,04	80	66	1,21	22	89	0,25	42	24	1,75	62	80	0,78	59	66	0,89	26	107	0,24
Casco Histórico de Vallecas	162	230	0,70	114	374	0,30	34	319	0,11	16	148	0,11	171	223	0,77	98	293	0,33	28	166	0,17	24	50	0,48
Orcasur	11	5	2,20	40	32	1,25	18	50	0,36	63	96	0,66	7	8	0,88	33	40	0,83	29	63	0,46	53	159	0,33
Zofío	61	26	2,35	17	90	0,19	13	32	0,41	5	24	0,21	43	70	0,61	20	191	0,10	5	31	0,16	5	30	0,17
Pradolongo	111	50	2,22	37	101	0,37	66	39	1,69	24	72	0,33	87	96	0,91	39	131	0,30	53	64	0,83	14	102	0,14
Abrantes	79	57	1,39	53	207	0,26	27	146	0,18	23	68	0,34	64	97	0,66	58	366	0,16	29	163	0,18	10	134	0,07
Puerta Bonita	128	132	0,97	82	307	0,27	43	131	0,33	20	117	0,17	151	161	0,94	96	413	0,23	20	235	0,09	23	191	0,12
Almendrales	99	48	2,06	54	114	0,47	23	52	0,44	35	33	1,06	72	72	1,00	53	203	0,26	24	90	0,27	24	80	0,30
Santa Eugenia	16	21	0,76	14	45	0,31	105	225	0,47	33	131	0,25	19	12	1,58	29	39	0,74	119	160	0,74	35	71	0,49
Vista Alegre	225	148	1,52	190	387	0,49	68	168	0,40	28	80	0,35	189	187	1,01	156	607	0,26	49	269	0,18	20	131	0,15
Entrevías	122	35	3,49	113	225	0,50	116	196	0,59	58	157	0,37	100	52	1,92	72	308	0,23	65	191	0,34	58	141	0,41
Águilas	120	17	7,06	216	225	0,96	68	106	0,64	20	44	0,45	96	40	2,40	216	319	0,68	72	122	0,59	18	65	0,28
Palomeras Sureste	116	55	2,11	63	236	0,27	124	127	0,98	107	244	0,44	143	124	1,15	57	296	0,19	122	195	0,63	69	191	0,36
Palomeras Bajas	122	78	1,56	116	168	0,69	55	90	0,61	21	91	0,23	145	150	0,97	112	282	0,40	51	121	0,42	23	144	0,16
Legazpi	51	58	0,88	26	37	0,70	54	67	0,81	46	92	0,50	18	91	0,20	27	70	0,39	24	127	0,19	27	192	0,14
Moscardó	62	45	1,38	109	161	0,68	42	148	0,28	28	71	0,39	77	99	0,78	69	277	0,25	21	261	0,08	7	102	0,07
Opañel	104	65	1,60	106	277	0,38	42	117	0,36	34	48	0,71	131	163	0,80	72	421	0,17	33	178	0,19	11	125	0,09
San Diego	454	355	1,28	159	449	0,35	47	213	0,22	19	102	0,19	359	533	0,67	99	456	0,22	37	283	0,13	23	88	0,26
Portazgo	104	80	1,30	81	185	0,44	67	81	0,83	21	54	0,39	104	94	1,11	93	159	0,58	63	75	0,84	24	84	0,29

Barrio	2018												2019											
	< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*			< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*		
	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W
Comillas	53	76	0,70	50	132	0,38	36	112	0,32	6	182	0,03	66	106	0,62	49	216	0,23	23	189	0,12	5	298	0,02
Campamento	87	57	1,53	45	75	0,60	36	31	1,16	39	37	1,05	87	56	1,55	49	114	0,43	26	38	0,68	23	53	0,43
Chopera	84	163	0,52	62	151	0,41	55	111	0,50	28	89	0,31	54	228	0,24	8	235	0,03	23	195	0,12	9	192	0,05
Delicias	120	195	0,62	75	58	1,29	43	93	0,46	47	180	0,26	102	365	0,28	61	151	0,40	45	172	0,26	81	230	0,35
Aluche	126	50	2,52	369	401	0,92	146	209	0,70	51	139	0,37	127	44	2,89	316	547	0,58	62	225	0,28	36	169	0,21
Pavones	20	6	3,33	13	28	0,46	32	27	1,19	7	19	0,37	27	3	9,00	5	52	0,10	21	30	0,70	4	22	0,18
Valderrivas	4	4	1,00	35	4	8,75	23	29	0,79	38	90	0,42	2	2	1,00	16	12	1,33	15	51	0,29	35	119	0,29
San Isidro	219	98	2,23	138	322	0,43	61	118	0,52	33	145	0,23	174	247	0,70	123	460	0,27	50	290	0,17	31	149	0,21
Valdebernardo	1	52	0,02	14	167	0,08	36	13	2,77	13	21	0,62	1	59	0,02	16	271	0,06	19	33	0,58	9	20	0,45
Numancia	353	244	1,45	145	328	0,44	54	111	0,49	56	96	0,58	319	420	0,76	122	431	0,28	49	167	0,29	50	160	0,31
Adelfas	58	38	1,53	42	64	0,66	36	74	0,49	23	113	0,20	136	91	1,49	68	113	0,60	53	127	0,42	53	125	0,42
Acacias	25	131	0,19	12	121	0,10	11	90	0,12	13	291	0,04	17	302	0,06	4	148	0,03	7	157	0,04	4	291	0,01
Pacífico	118	173	0,68	84	100	0,84	86	123	0,70	84	263	0,32	102	158	0,65	67	108	0,62	55	181	0,30	49	457	0,11
Palos de la Frontera	163	263	0,62	71	145	0,49	44	88	0,50	58	152	0,38	158	399	0,40	67	192	0,35	48	146	0,33	57	317	0,18
Fontarrón	63	14	4,50	89	113	0,79	20	62	0,32	17	29	0,59	77	22	3,50	65	183	0,36	22	97	0,23	8	40	0,20
Atocha	1	2	0,50	0	10	0,00	1	0	-	2	4	0,50	3	10	0,30	0	0	-	0	0	-	3	0	-
Los Cármenes	36	61	0,59	64	87	0,74	46	74	0,62	19	58	0,33	46	28	1,64	52	150	0,35	32	107	0,30	16	64	0,25
Vinateros	53	35	1,51	55	92	0,60	38	78	0,49	7	7	1,00	38	15	2,53	48	92	0,52	41	91	0,45	3	30	0,10
Lucero	155	86	1,80	169	188	0,90	74	149	0,50	28	93	0,30	130	158	0,82	159	378	0,42	62	191	0,32	26	148	0,18
Horcajo	5	0	-	7	5	1,40	3	4	0,75	8	14	0,57	0	0	-	3	0	-	9	7	1,29	3	36	0,08
Imperial	37	90	0,41	42	82	0,51	43	79	0,54	38	116	0,33	38	118	0,32	27	136	0,20	23	165	0,14	36	213	0,17
Embajadores	700	1.293	0,54	157	361	0,43	75	303	0,25	94	398	0,24	611	1.493	0,41	122	556	0,22	83	406	0,20	64	510	0,13
Marroquina	57	36	1,58	18	26	0,69	52	41	1,27	20	56	0,36	43	26	1,65	12	20	0,60	76	106	0,72	26	62	0,42
Niño Jesús	21	14	1,50	17	19	0,89	32	63	0,51	58	192	0,30	15	25	0,60	16	20	0,80	13	55	0,24	59	187	0,32
Puerta del Ángel	249	235	1,06	186	379	0,49	66	172	0,38	50	87	0,57	188	322	0,58	173	635	0,27	46	228	0,20	26	132	0,20
Media Legua	48	9	5,33	45	52	0,87	29	28	1,04	22	39	0,56	53	14	3,79	30	57	0,53	26	64	0,41	24	109	0,22
Cortes	96	171	0,56	37	118	0,31	27	57	0,47	68	296	0,23	91	195	0,47	49	107	0,46	21	80	0,26	37	426	0,09
Sol	33	78	0,42	19	68	0,28	22	94	0,23	40	330	0,12	22	220	0,10	19	144	0,13	14	94	0,15	29	330	0,09
Estrella	12	7	1,71	55	41	1,34	27	57	0,47	96	217	0,44	16	11	1,45	28	26	1,08	22	99	0,22	70	240	0,29
Los Jerónimos	11	11	1,00	13	20	0,65	5	22	0,23	87	178	0,49	8	13	0,62	6	11	0,55	4	25	0,16	62	242	0,26
Ibiza	71	108	0,66	46	59	0,78	64	99	0,65	152	335	0,45	66	132	0,50	58	149	0,39	62	180	0,34	84	419	0,20

Barrio	2018												2019											
	< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*			< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*		
	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W
Palacio	115	313	0,37	40	195	0,21	41	160	0,26	31	382	0,08	115	513	0,22	46	353	0,13	22	223	0,10	30	558	0,05
Arcos	116	69	1,68	37	61	0,61	78	99	0,79	9	41	0,22	105	111	0,95	22	94	0,23	74	134	0,55	5	83	0,06
Goya	154	387	0,40	62	160	0,39	48	151	0,32	156	647	0,24	117	362	0,32	61	381	0,16	41	281	0,15	105	909	0,12
Amposta	119	133	0,89	13	40	0,33	16	12	1,33	3	24	0,13	108	148	0,73	12	60	0,20	14	18	0,78	2	3	0,67
Justicia	142	295	0,48	67	182	0,37	56	83	0,67	180	381	0,47	131	371	0,35	55	163	0,34	50	176	0,28	123	693	0,18
Recoletos	48	40	1,20	23	94	0,24	24	65	0,37	153	456	0,34	25	70	0,36	16	95	0,17	10	104	0,10	110	714	0,15
Universidad	318	916	0,35	123	310	0,40	66	183	0,36	102	518	0,20	278	1.236	0,22	135	555	0,24	69	365	0,19	130	976	0,13
Fuente del Berro	118	274	0,43	65	93	0,70	52	78	0,67	66	230	0,29	106	252	0,42	54	221	0,24	37	166	0,22	42	298	0,14
Ventas	264	205	1,29	248	359	0,69	31	132	0,23	39	69	0,57	248	270	0,92	209	559	0,37	38	262	0,15	24	115	0,21
Hellín	41	65	0,63	39	51	0,76	19	15	1,27	45	18	2,50	45	89	0,51	32	69	0,46	11	12	0,92	42	45	0,93
Argüelles	103	134	0,77	79	122	0,65	46	102	0,45	133	439	0,30	65	155	0,42	37	159	0,23	25	136	0,18	108	593	0,18
Lista	92	157	0,59	98	132	0,74	33	128	0,26	113	439	0,26	67	280	0,24	32	257	0,12	24	147	0,16	59	715	0,08
Pueblo Nuevo	305	216	1,41	244	492	0,50	96	271	0,35	49	156	0,31	285	305	0,93	193	646	0,30	74	317	0,23	26	226	0,12
Castellana	51	55	0,93	25	71	0,35	20	46	0,43	121	448	0,27	42	108	0,39	16	112	0,14	13	126	0,10	103	510	0,20
Almagro	51	100	0,51	36	78	0,46	28	20	1,40	138	462	0,30	51	233	0,22	30	117	0,26	17	127	0,13	147	849	0,17
Trafalgar	161	293	0,55	64	123	0,52	62	102	0,61	109	298	0,37	169	432	0,39	54	216	0,25	40	181	0,22	87	526	0,17
Arapiles	159	250	0,64	82	187	0,44	43	92	0,47	82	275	0,30	134	359	0,37	60	205	0,29	36	130	0,28	63	337	0,19
Gaztambide	109	127	0,86	47	72	0,65	39	86	0,45	118	473	0,25	90	223	0,40	51	134	0,38	40	163	0,25	92	639	0,14
Quintana	125	110	1,14	130	182	0,71	50	100	0,50	14	88	0,16	136	217	0,63	90	240	0,38	22	135	0,16	16	104	0,15
Simancas	227	194	1,17	81	156	0,52	77	96	0,80	113	141	0,80	194	327	0,59	80	181	0,44	43	164	0,26	63	224	0,28
Guindalera	205	315	0,65	89	188	0,47	91	202	0,45	208	576	0,36	147	393	0,37	82	290	0,28	77	379	0,20	138	880	0,16
La Concepción	116	59	1,97	93	138	0,67	27	111	0,24	25	53	0,47	106	93	1,14	74	251	0,29	16	153	0,10	9	79	0,11
Ríos Rosas	144	106	1,36	64	85	0,75	49	67	0,73	137	303	0,45	128	223	0,57	54	243	0,22	47	161	0,29	100	599	0,17
Vallehermoso	29	22	1,32	31	31	1,00	38	37	1,03	111	221	0,50	40	41	0,98	49	33	1,48	23	93	0,25	90	454	0,20
San Pascual	75	41	1,83	32	87	0,37	25	33	0,76	51	90	0,57	66	67	0,99	16	101	0,16	32	53	0,60	28	182	0,15
Canillejas	82	27	3,04	94	169	0,56	53	93	0,57	37	115	0,32	101	65	1,55	121	244	0,50	41	157	0,26	49	167	0,29
Rosas	24	23	1,04	34	24	1,42	61	91	0,67	51	140	0,36	20	13	1,54	31	30	1,03	48	128	0,38	38	189	0,20
El Salvador	18	11	1,64	15	43	0,35	23	33	0,70	47	94	0,50	20	6	3,33	12	66	0,18	21	31	0,68	32	182	0,18
Rejas	22	53	0,42	23	44	0,52	72	98	0,73	54	137	0,39	226	92	2,46	43	77	0,56	55	176	0,31	47	202	0,23
Casa de Campo	18	14	1,29	40	41	0,98	22	44	0,50	39	90	0,43	21	17	1,24	26	58	0,45	30	74	0,41	31	73	0,42
El Viso	18	28	0,64	31	22	1,41	19	41	0,46	122	320	0,38	27	32	0,84	25	30	0,83	28	53	0,53	93	520	0,18

Barrio	2018												2019											
	< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*			< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*		
	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W
Ciudad Jardín	116	84	1,38	59	61	0,97	48	62	0,77	58	167	0,35	86	129	0,67	56	109	0,51	43	190	0,23	38	226	0,17
Prosperidad	122	195	0,63	125	175	0,71	101	163	0,62	71	260	0,27	134	333	0,40	121	448	0,27	66	408	0,16	59	598	0,10
San Juan Bautista	15	6	2,50	19	26	0,73	36	43	0,84	47	83	0,57	24	30	0,80	29	50	0,58	31	56	0,55	41	135	0,30
Cuatro Caminos	193	202	0,96	88	106	0,83	54	100	0,54	149	320	0,47	188	351	0,54	92	239	0,38	73	220	0,33	158	649	0,24
Palomas	0	0	-	3	0	-	13	8	1,63	46	119	0,39	0	0	-	2	10	0,20	7	20	0,35	44	299	0,15
Bellas Vistas	289	375	0,77	98	159	0,62	39	87	0,45	37	139	0,27	196	430	0,46	82	267	0,31	29	170	0,17	30	126	0,24
Hispanoamérica	52	42	1,24	78	69	1,13	36	127	0,28	131	427	0,31	50	46	1,09	72	124	0,58	56	165	0,34	137	477	0,29
Piovera	7	4	1,75	12	7	1,71	21	15	1,40	106	299	0,35	9	15	0,60	12	11	1,09	13	44	0,30	70	332	0,21
Colina	17	6	2,83	23	31	0,74	21	19	1,11	23	47	0,49	21	23	0,91	18	21	0,86	12	30	0,40	26	55	0,47
Berruguete	172	292	0,59	68	169	0,40	40	106	0,38	23	103	0,22	199	326	0,61	85	334	0,25	35	143	0,24	23	150	0,15
Alameda de Osuna	12	3	4,00	43	10	4,30	30	43	0,70	131	130	1,01	66	1	66,00	84	7	12,00	52	88	0,59	86	207	0,42
Castillejos	101	91	1,11	62	64	0,97	37	129	0,29	82	267	0,31	99	133	0,74	66	167	0,40	38	210	0,18	69	323	0,21
Aravaca	103	17	6,06	143	30	4,77	55	56	0,98	220	397	0,55	26	52	0,50	14	116	0,12	22	60	0,37	143	643	0,22
Atalaya	3	0	-	13	8	1,63	1	9	0,11	6	23	0,26	6	1	6,00	12	12	1,00	3	28	0,11	2	33	0,06
Nueva España	42	50	0,84	44	58	0,76	16	94	0,17	100	404	0,25	40	56	0,71	46	98	0,47	31	93	0,33	100	519	0,19
Canillas	80	61	1,31	118	115	1,03	101	77	1,31	61	203	0,30	66	77	0,86	105	191	0,55	87	175	0,50	41	315	0,13
Corralejos	11	4	2,75	11	4	2,75	18	27	0,67	66	114	0,58	8	6	1,33	4	15	0,27	19	38	0,50	33	154	0,21
Valdezarza	74	89	0,83	142	164	0,87	42	66	0,64	68	97	0,70	63	138	0,46	125	219	0,57	41	99	0,41	65	72	0,90
Valdeacederas	233	170	1,37	90	296	0,30	34	138	0,25	16	83	0,19	222	240	0,93	113	372	0,30	33	234	0,14	28	146	0,19
Almenara	210	96	2,19	182	92	1,98	90	29	3,10	88	150	0,59	89	73	1,22	64	89	0,72	59	75	0,79	78	189	0,41
Valdemarín	8	17	0,47	7	10	0,70	9	8	1,13	85	143	0,59	1	8	0,13	1	8	0,13	4	10	0,40	56	226	0,25
Casco Histórico de Barajas	30	21	1,43	34	35	0,97	25	24	1,04	10	37	0,27	26	74	0,35	26	58	0,45	14	42	0,33	6	53	0,11
Ciudad Universitaria	35	35	1,00	22	47	0,47	16	19	0,84	73	151	0,48	25	46	0,54	19	41	0,46	11	33	0,33	46	237	0,19
El Plantío	2	0	-	0	2	0,00	5	6	0,83	17	80	0,21	0	1	0,00	0	2	0,00	1	0	-	17	114	0,15
Pinar del Rey	288	85	3,39	166	226	0,73	110	119	0,92	59	166	0,36	240	164	1,46	160	287	0,56	154	200	0,77	74	189	0,39
Apóstol Santiago	12	7	1,71	49	68	0,72	26	35	0,74	44	58	0,76	12	7	1,71	49	129	0,38	29	63	0,46	30	99	0,30
Castilla	19	12	1,58	32	55	0,58	25	37	0,68	66	251	0,26	10	47	0,21	37	147	0,25	20	51	0,39	58	326	0,18
Costillares	12	8	1,50	26	18	1,44	35	38	0,92	90	228	0,39	19	25	0,76	24	50	0,48	33	104	0,32	83	427	0,19
Pilar	140	139	1,01	26	170	0,15	7	29	0,24	36	90	0,40	189	203	0,93	43	206	0,21	26	46	0,57	52	124	0,42
Peñagrande	83	69	1,20	125	110	1,14	89	80	1,11	204	257	0,79	86	60	1,43	105	135	0,78	55	133	0,41	101	473	0,21

Barrio	2018												2019											
	< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*			< 60 *m2*			60 – 80 *m2*			80 – 100 *m2*			> 100 *m2*		
	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W	O	P	W
La Paz	5	14	0,36	33	24	1,38	53	35	1,51	107	201	0,53	11	5	2,20	20	48	0,42	29	46	0,63	92	320	0,29
Fuentelarreina	1	0	-	3	5	0,60	2	4	0,50	31	57	0,54	1	0	-	1	2	0,50	2	5	0,40	24	94	0,26
Timón	17	16	1,06	87	30	2,90	33	15	2,20	18	35	0,51	42	23	1,83	31	31	1,00	30	38	0,79	22	76	0,29
Aeropuerto	17	25	0,68	2	7	0,29	0	0	-	0	3	0,00	10	24	0,42	2	7	0,29	1	3	0,33	1	0	-
Valdefuentes	149	20	7,45	217	51	4,25	73	96	0,76	987	407	2,43	87	50	1,74	107	79	1,35	79	129	0,61	779	644	1,21
Mirasierra	14	20	0,70	33	3	11,00	69	29	2,38	244	277	0,88	15	5	3,00	22	7	3,14	57	53	1,08	366	259	1,41
Valverde	151	68	2,22	214	167	1,28	138	172	0,80	171	215	0,80	138	131	1,05	144	300	0,48	129	221	0,58	195	434	0,45
El Goloso	9	5	1,80	22	12	1,83	69	19	3,63	146	125	1,17	15	20	0,75	30	15	2,00	35	26	1,35	91	177	0,51
El Pardo	6	1	6,00	16	4	4,00	13	6	2,17	10	12	0,83	1	0	-	3	7	0,43	8	4	2,00	6	14	0,43
Casco histórico de Vicálvaro	188	28	6,71	87	68	1,28	27	38	0,71	9	9	1,00	156	48	3,25	74	145	0,51	21	46	0,46	22	19	1,16
El Cañaveral	0	0	-	39	4	9,75	63	11	5,73	192	42	4,57	1	5	0,20	36	10	3,60	173	24	7,21	358	59	6,07

Fuente: Elaboración propia basada en Ayuntamiento de Madrid (2020).

Tabla 4.
Resumen de los modelos utilizados por cada uno de los autores que estimaron el TOM.

Autores	Modelo utilizado
Haurin (1988)	Regresión logit
Forgey *et al.* (1996)	Regresión logit
Genesove y Mayer (1997)	Análisis de supervivencia: regresión de Cox
Anglin ***et al.*** (2001)	Análisis de supervivencia: Weibull
Khezr (2015)	Regresión logit
Li (2015)	Análisis de supervivencia: Weibull
Cirman ***et al.*** (2015)	Análisis de supervivencia: Weibull
Bhuiyan y Hasan (2016)	Análisis de supervivencia: regresión de Cox
Scofield y Devaney (2017)	Regresión probit
Bich *et al.* (2020)	Análisis de supervivencia: regresión de Cox

Fuente: elaboración propia.

Para contrastar este proceso metodológico de carácter general, asumiendo que cada zona geográfica tendrá unas particularidades, si bien los preceptos generales de comportamiento serán coincidentes, el análisis se ha particularizado sobre la capital de España, Madrid, sencillamente para testear los datos en una localidad dotada de un mercado lo suficientemente grande, líquido y diverso. Madrid es, por tanto, una de las ciudades más representativas de España y así las conclusiones obtenidas de su estudio tienen ese carácter general y útil; es el municipio con más habitantes (INE, 2023b) y el que tiene mayor número de transacciones inmobiliarias (Ministerio de Transportes y Movilidad Sostenible, 2023). Los datos han sido recopilados a través de las publicaciones de inmuebles de segunda mano llevadas a cabo en el portal inmobiliario líder de España, Idealista, durante los años 2018 y 2019 –evitando los años más recientes, 2020 y 2021, extremadamente anormales por la pandemia COVID-19–, donde el volumen total que se dispone del portal está formado por 179.533 inmuebles (96% de ellos son pisos), donde un 85% fueron vendidos (86% en pisos y 74% en chalés), como se puede apreciar por barrios en la Tabla 3 (153.077, que corresponden a 72.321 según las fuentes oficiales). Además, el tiempo de venta medio son de seis y 11 semanas (para chalés y pisos, respectivamente) mientras que se observan alrededor de 25.000 pisos publicados conjuntamente y 1.600 chalés. Por todo esto,

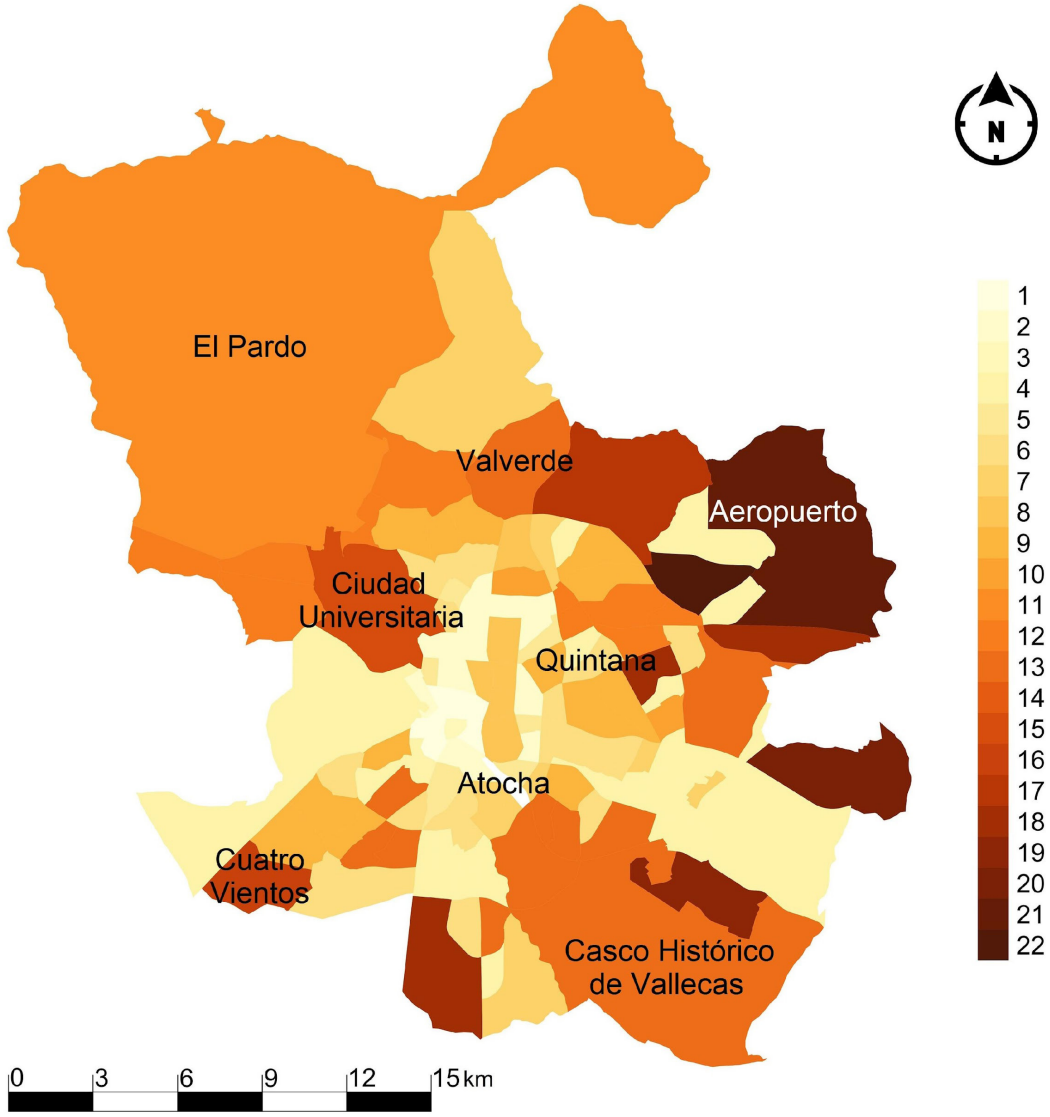
y dada la naturaleza específica de cada tipo de vivienda, la modelización de cada una se ha realizado de forma separada.

Como se ha comentado anteriormente, se había de elaborar los indicadores de sobreprecio, siendo necesario calcular un precio teórico para los inmuebles –de cara a la obtención del DOP del inmueble en cuestión–, por lo que, mediante técnicas de *machine learning,* y con objeto de tratar de normalizar la variable y minimizar su variabilidad, se estimó el logaritmo del precio para aquellos inmuebles que fueron vendidos. Una vez obtenido este valor, el cálculo del DOP fue inmediato.

Así mismo, se establecieron una serie de puntuaciones para definir cada uno de los 135 barrios considerados de Madrid –a los 131 oficiales se añaden cuatro adicionales, que, debido a ser más novedosos e identificables por sus características propias, se han mantenido aparte por venir así ya recogidos en la muestra: Pau de Carabanchel, Arroyo del Fresno, Las Tablas y Sanchinarro–, lo que dio lugar a otro sobreprecio: el relativo al barrio. Como los atributos externos de los que se disponía eran muy numerosos, se decidió crear unos ejes que representaran de un modo notorio la localización de los inmuebles (y así su probabilidad de venta) a través de tres bloques: variables económicas (ingresos, esfuerzo hipotecario/alquiler, precios medianos de venta/alquiler, relación entre alquileres y ventas, entre otros), puntos de interés (*point of interest,* PoI, como restauración, puntos de interés turístico, establecimientos comerciales, paradas de transporte o centros de salud) y otros (nacionalidad, tipos de inmueble y antigüedad de las viviendas), utilizando para las dos primeras un análisis de componentes principales (método de reducción de variables, mediante el cual se genera una serie de componentes incorrelacionados que expliquen la mayor parte de la varianza total) y para las terceras un análisis de correspondencias (a partir de una tabla de contingencia formada por diferentes variables nominales y ordinales, permite representar sus patrones a través de una serie de valores). Estas puntuaciones se pueden representar a través de un análisis clúster, generando una agrupación formada por 22 conjuntos, que, aunque no se utilizará en el desarrollo de los modelos (sino los valores de éstas), se mostrará a nivel informativo en la Figura 1 con objeto de corroborar que lo generado cobra sentido, donde las diferencias cromáticas representan cada uno de los grupos (Tabla 5).

Tras ello, ya se dispone por completo de todas las variables a utilizar en la modelización. Se presenta en la Tabla 6 un listado de las variables independientes que marcarán el desarrollo de la transacción de un inmueble, agrupadas en cuatro bloques diferenciados, mientras que en la Tabla 7 se recogen las dos variables dependientes propias de este tipo de modelos.

Figura 1.
Visualización del clúster k-medoids según los atributos externos de los barrios de Madrid.



Fuente: elaboración propia.

Tabla 5.
Agrupación visual obtenida de los diferentes barrios de Madrid.

Grupo	Barrio
1	Palacio, Cortes, Justicia y Universidad
2	Embajadores, Palos de la Frontera, Niño Jesús, Goya, Lista, Ciudad Jardín, Hispanoamérica, Cuatro Caminos, Castillejos, Arapiles, Trafalgar, Ríos Rosas, Vallehermoso y Argüelles
3	Sol
4	Imperial, Casa de Campo, Los Cármenes, Campamento, Orcasitas, Orcasur, San Fermín, Zofío, Pradolongo, Pavones, Fontarrón, San Pascual, Apóstol Santiago, San Cristóbal, Santa Eugenia, Casco Histórico de Vicálvaro, Valderrivas, Hellín, Amposta, Alameda de Osuna, Casco Histórico de Barajas y Timón
5	Acacias, Pacífico, Ibiza, Prosperidad, Berruguete, Gaztambide y Comillas
6	Chopera, Estrella, Fuente del Berro, Bellas Vistas, Almenara, Valdeacederas, Valdezarza, Lucero, Opañel, Buenavista, Almendrales, Moscardó, Portazgo, Marroquina, Media Legua, Vinateros, Quintana, La Concepción, Ángeles y Canillejas
7	Legazpi, Delicias, Adelfas, El Goloso, Pau de Carabanchel, Horcajo, Atalaya, Costillares, Atocha, Butarque y Valdebernardo
8	Los Jerónimos, Recoletos, Castellana, El Viso, Castilla y Almagro
9	Guindalera, Peñagrande, Pilar, La Paz, Puerta del Ángel, Aluche, Águilas, Vista Alegre, Numancia, Ventas, Pueblo Nuevo, Canillas y Pinar del Rey
10	Nueva España y Arcos
11	El Pardo
12	Fuentelarreina, Mirasierra, Arroyo del Fresno, Valdemarín, El Plantío, Aravaca, San Juan Bautista, Colina, Palomas, Piovera y El Salvador
13	Valverde, San Isidro, Puerta Bonita, Abrantes, Entrevías, San Diego, Palomeras Bajas, Palomeras sureste, Los Rosales, Casco Histórico de Vallecas y Rosas
14	Las Tablas y Sanchinarro
15	Ciudad Universitaria
16	Cuatro Vientos
17	Valdefuentes
18	Villaverde Alto - Casco Histórico de Villaverde, Simancas y Rejas
19	Ensanche de Vallecas
20	El Cañaveral
21	Aeropuerto
22	Corralejos

Fuente: elaboración propia.

Tabla 6.
Variables independientes utilizadas en la estimación de la probabilidad de venta de un inmueble, recogiendo tanto las de los pisos, como las de los chalés.

Variable	Descripción	Tipo
Atributos internos: caracterizan la vivienda en sí misma		
area	Superficie de la vivienda, en	Numérica
habitaciones	Número de habitaciones de la vivienda	Numérica
baños	Número de baños de la vivienda	Numérica
antiguedad	Antigüedad de la vivienda (años que han pasado desde su fecha de construcción)	Numérica
orientacion	Orientación de la vivienda (Norte, Sur, Este y Oeste, más sus combinaciones)	Categórica
exterior	Muestra si la vivienda es exterior o interior	Dicotómica
numero_variaciones_precio	Número de variaciones de precio sufridas desde la publicación del anuncio	Numérica
variacion_precio	Variación porcentual de precio sufrida desde la publicación del anuncio	Numérica
ascensor	Indica la presencia o no de ascensor	Dicotómica
garaje	Indica la presencia o no de garaje	Dicotómica
trastero	Indica la presencia o no de trastero	Dicotómica
piscina	Indica la presencia o no de piscina	Dicotómica
jardin	Indica la presencia o no de jardín	Dicotómica
terraza	Indica la presencia o no de terraza	Dicotómica
balcon	Indica la presencia o no de balcón	Dicotómica
portero	Indica la presencia o no de portero	Dicotómica
atico	Indica, en el caso de no ser un chalé, si es un ático o no	Dicotómica
duplex	Indica, en el caso de no ser un chalé, si es un dúplex o no	Dicotómica
armarios_empotrados	Indica la presencia o no de armarios empotrados	Dicotómica
aire_acondicionado	Indica la presencia o no de aire acondicionado	Dicotómica
anunciado_empresa	Indica si el inmueble ha sido anunciado por empresa (o de forma particular)	Dicotómica
chale_adosado	Indica, en caso de ser un chalé, si es adosado o no	Dicotómica
dimension_parcela	Indica, en caso de ser un chalé y disponerla, la superficie de la parcela, en	Numérica

Atributos externos: puntuaciones que marcan su localización		
riqueza_barrio	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, para determinar su riqueza económica	Numérica
alquiler_venta	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, que determina la relación entre los precios de alquiler y de venta de los inmuebles	Numérica
esfuerzo_alquiler	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, para mostrar el esfuerzo económico dedicado al alquiler de viviendas en él	Numérica
densidad_poblacion	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, para mostrar su densidad de población	Numérica
superficie_tipos_inmuebles	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, para mostrar la distribución de las superficies de los diferentes tipos de inmuebles (habitacional, comercial, oficina o local, entre otros)	Numérica
actividades_turisticas_culinarias	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, que determina sus puntos de interés turísticos y culinarios	Numérica
actividades_generales	Puntuación, del Análisis de Componentes Principales para el barrio de la vivienda, que determina su importancia para actividades de uso cotidiano (como paradas de transporte o supermercados)	Numérica
tipos_inmuebles	Puntuación, del Análisis de Correspondencias para el barrio de la vivienda, que hace referencia a la distribución de oficinas, zonas comerciales y locales dentro de él	Numérica
nacionalidad	Puntuación, del Análisis de Correspondencias para el barrio de la vivienda, que determina su distribución de nacionalidades	Numérica
antiguedad_viviendas	Puntuación, del Análisis de Correspondencias para el barrio de la vivienda, que determina la antigüedad de sus inmuebles	Numérica
Interés del comprador: datos semanales		
contactos	Número de contactos semanales recibidos por la vivienda	Numérica
visitas	Número de visitas semanales al anuncio de la vivienda	Numérica
apariciones	Número de apariciones semanales de la vivienda en búsquedas por los usuarios	Numérica
Precio: bruto y relativizado		
precio	Precio anunciado de la vivienda, en €	Numérica
dop_inmueble	DOP del precio anunciado de la vivienda, respecto a su precio teórico	Numérica
dop_barrio	DOP del precio anunciado de la vivienda, respecto al precio mediano de su barrio	Numérica

Fuente: elaboración propia.

Tabla 7.
Variables dependientes utilizadas en la estimación de la probabilidad de venta de un inmueble.

Variable	Descripción	Tipo
estado	Informa de la situación de la vivienda en una semana concreta: publicada para vender o ya vendida	Dicotómica
tiempo	Tiempo que tarda la vivienda en ser vendida, en semanas	Numérica

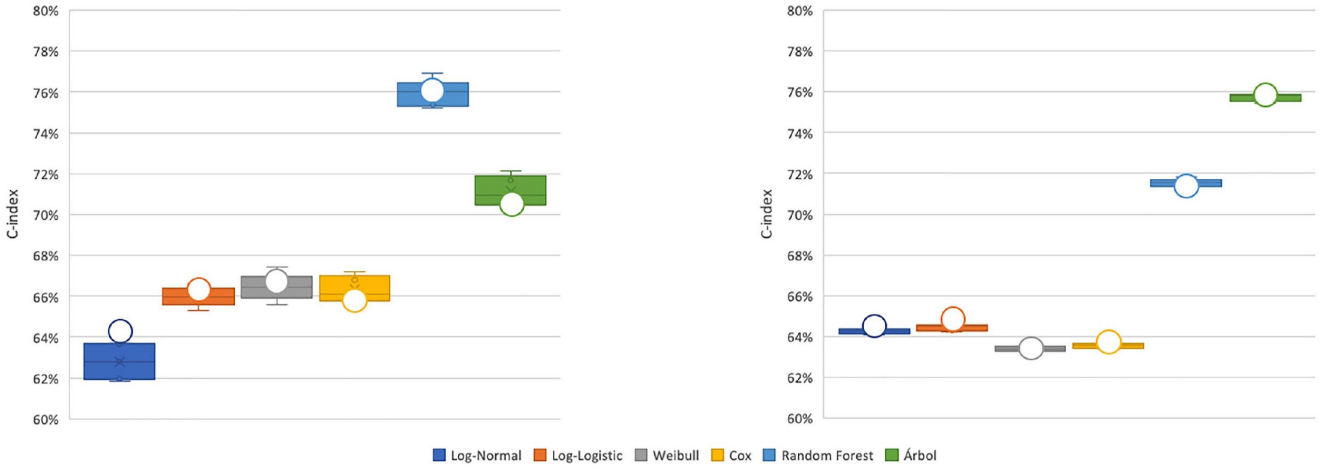
Fuente: elaboración propia.

Resultados

En la estimación del precio teórico de los inmuebles –uno de los objetivos secundarios, pero fundamentales de abordar en el estudio para así poder elaborar el indicador que marcase el DOP–, se obtuvieron métricas de R^2 que superaban el 94% (el modelo ganador fue el *random forest*) y 95% (***gradient boosting machine***) en chalés y pisos, respectivamente, sobre la muestra de test (con 20% de los registros). Cabe destacar que estos resultados son todavía mejores que los obtenidos previamente en los estudios comentados, donde ninguno superaba el 92%. La mejora obtenida en nuestro trabajo cobra aún mayor valor debido a la reponderación comentada, proceso no aplicado en los previos.

De cara a la consecución del objetivo principal del trabajo, se modelizaron diferentes funciones de supervivencia con los algoritmos comentados: paramétricos (log-normal, log-logistic y weibull, todos con AFT), semiparamétricos (Cox) y los provenientes del ***machine learning*** (***random forest*** y árboles). Se muestran en la Figura 2 los resultados obtenidos, donde el diagrama de cajas y bigotes representa el entrenamiento con CV, mientras el punto blanco muestra el resultado obtenido en test.

Figura 2.
Resultados obtenidos de chalés (a la izquierda) y pisos (derecha).



Fuente: elaboración propia.

Tras analizar los diferentes modelos, los mejores resultados obtenidos de los *C-index* (Harrell *et al.*, 1982) rondan el 76% y el 72% en test, para chalés y pisos respectivamente, siendo el mejor en ambos casos el *Random Forest*. A pesar de que la métrica del Árbol superaba en 4pb al *Random Forest,* su resultado en la aplicación dejó que desear al no proporcionar grandes diferencias numéricas a nivel inmueble, resultando a nivel agregado idénticas para varios distritos (el gran peso de la variable contactos es determinante en este caso), demostrando que este tipo de algoritmos consigue una capacidad predictiva superior a los tradicionalmente empleados en problemas de análisis de supervivencia (los modelos paramétricos y Cox arrojaron 66% en chalés y el 64% en pisos). Además, como fortalezas del *Random Forest,* se observan una distribución de los pesos más homogénea y un algoritmo común a ambos tipos de inmueble, a cambio de un elevado tiempo de procesamiento en comparación con el resto, por eso se seleccionará este algoritmo como modelo a considerar.

Los resultados obtenidos son notablemente destacables al compararlos con investigaciones previas. Aunque no es factible homogeneizar los datos, ya que se usan distintas métricas en cada uno de los estudios, en la Tabla 2 se aprecian valores relativamente inferiores para ellos.

En cuanto a la importancia de las variables en los modelos de *Random Forest* obtenidos, véase la Figura 3, donde se significan como las más determinantes en ambos modelos aquellas relativas al interés suscitado por el comprador (apariciones en búsquedas, visitas del inmueble en concreto y/o contactos con el vendedor así lo muestran), junto con el DOP del inmueble (para diferenciar las viviendas que están en valor de mercado de las que no). Tras ellas, y de manera particular para la categoría de chalés, les siguen el tamaño de la parcela, DOP del barrio e indicadores de localización (permite comparar ese inmueble con otro de la misma zona),

área y precio; para pisos, el modelo identifica también como variables significativas que el tipo anunciante sea empresa (posiblemente se persigue dar salida a inmuebles que proceden de embargos), variación desde inicio (cuántas se han realizado y en qué porcentaje marcarán la existencia de disposición a negociar por parte del vendedor) e indicadores de localización.

Atendiendo a cada una de las variables explicativas del modelo, las principales características del inmueble que marcarán el tiempo de la compraventa del mismo son las siguientes:

1) Internas. Identifican la vivienda en sí misma. De modo general, inmuebles pequeños, antiguos, anunciados por empresas, con la ausencia de características adicionales y/o sin ajustes de precio tienen relación directa con TOM. En cuanto a la orientación, los esquinados son los más difíciles de vender.

2) Externas, definidas por su localización. Aquellas zonas más exclusivas tienen menor probabilidad de venta: barrios más poblados o alejados de la urbe, con mayor riqueza económica, cercanas a puntos de interés no turísticos/culinarios y/o con una concentración mayor de población inmigrante.

3) Interés del comprador, representado por las apariciones en búsquedas, visitas del inmueble y/o contactos con el vendedor, indica una mayor probabilidad de venta.

4) Precio: el precio informado guarda una relación directa con TOM, así como el DOP del barrio para pisos. Por su parte, en cuanto al DOP propio del inmueble en cuestión, las viviendas que más pronto se venden son aquellas que tienen un precio más cercano a su valor teórico.

Así mismo, en la Figura 4 se muestran las curvas de supervivencia (probabilidad de no venta) de los 21 distritos de Madrid que agrupan sus 131 barrios, agregando los inmuebles que pertenecen a ellos y tomando sus valores medianos. Como se puede observar, el rango de las probabilidades en los chalés es más amplio, existiendo diferencias importantes entre los distritos (en Villaverde, por ejemplo, únicamente quedan sin vender en las diez primeras semanas un 10% de los inmuebles vs. 70% en Hortaleza y Moncloa-Aravaca), situándose por encima de un 15% de no venta en un año en tres distritos (Hortaleza, Chamartín y Moncloa-Aravaca), frente a un 1,6 - 3,6% de rango de no venta en los pisos.

Con el último valor anterior (probabilidad de no venta en el siguiente año desde la publicación en el portal), se puede constituir un mapa de calor de cada uno de los distritos para visualizar los valores de un modo geoespacial como la Figura 5, de modo que los tonos rojos presentan mayor probabilidad de venta, y los verdes, menor.

Discusión

En primer lugar, y como se ha comentado en la sección de Metodología, la modelización se realizó por separado entre pisos y chalés. Esto se apoyó en el hecho de que sus características, así como la motivación del comprador, son totalmente diferentes: quien desea un piso, busca estar dentro de la urbe, tener al alcance todos los servicios, además de no subir y bajar escaleras continuamente. Mientras que un chalé se suele adquirir por personas que priorizan la tranquilidad y el espacio por encima de todo, a menudo con una renta superior. En cualquier caso, este aspecto fue corroborado mediante el test ***logrank*** (Mantel, 1966) y las curvas de supervivencia de Kaplan-Meier de ambos, demostrando que los pisos se venden más rápido que los chalés como se puede apreciar en la Figura 6.

Mediante algoritmos de supervivencia, los más empleados en este tipo de problemas como se puede observar en la recopilación literaria de la Tabla 4, y yendo un paso más allá debido a la incorporación en ellos de técnicas de *machine learning*, se ha corroborado que los pisos más pequeños y antiguos son los que más rápido se venden, como afirmaban Forgey *et al.* (1996), aunque también se venden rápido aquellos con ausencia de atributo, lo cual contradice lo expuesto por Anglin *et al.* (2001). La presencia de características como piscina, jardín o garaje será deseada por los compradores en igualdad del resto de condiciones, pero la diferencia de precio es significativa en el presente estudio (por ejemplo, el precio mediano de venta en los pisos áticos o con piscina supera en más de un 80% a los que no disponen de estas propiedades) y quizás no en el de Anglin *et al.* (2001) –análisis, por cierto, llevado a cabo en un periodo económico bastante diferente al especificado aquí, pues data de 1997 y la muestra tuvo únicamente 3.874 registros sin re-ponderar– por lo que, en verdad, la motivación económica parece tener un peso significativo en la elección del comprador, decantándose por aquellos más asequibles, como ya afirmaban McCall y Lippman (1984). Así mismo, y aunque no se incluye directamente en este estudio, actualmente puede también influir otra motivación: la relacionada con las viviendas de uso turístico. En los últimos tiempos, éstas han experimentado un gran auge, por lo que aquellos inmuebles más baratos, mediante su alquiler, permitirán recuperar la inversión antes y la rentabilidad obtenida será mayor.

A nivel localización, ya Cirman *et al.* (2015) observaban diferencias, donde las zonas céntricas mostraban mayor tiempo de venta. Y en este estudio se ha confirmado esa tesis: las zonas de mayor riqueza (Norte y Centro de Madrid, contrastado en la Figura 5) y/o no del todo cercanas a PoIs turísticos/culinarios tienen menor probabilidad de venta. Consideramos el primer punto basados en el precio, barrera de entrada para gran parte de la población residente: son zonas que tienen mucha demanda y donde el suelo es escaso. El segundo, parece también sustentarse en el turismo: resulta preferible en un viaje alojarse en la zona donde se va a disfrutar del tiempo.

Por su parte, la relación con el DOP no es trivial, y las diferencias parecen reflejarse por la distancia a su precio teórico: con menor probabilidad de venta para aquellos más alejados de él, acentuándose más en los de elevado infra precio que en los de gran sobreprecio: si el precio es muy elevado, convendría esperar a que descienda; y si es muy bajo, quizá no interese comprarlo. Esto, debido a algo interno del inmueble (por ejemplo, que la casa necesite una reforma profunda, ya porque sea algo externo, como que la localización concreta así lo desaconseje). Por ello, se ha verificado la relación directa con el TOM como declaraban Anglin *et al.* (2001) y no inversa como lo hacían Bich *et al.* (2020), a pesar de que Madrid es una ciudad que verifica la escasa oferta que tenían como hipótesis en su estudio (si bien es cierto que tampoco había re-ponderación en su estudio y una muestra formada por 448 inmuebles). En cualquier caso, lo idóneo resultará establecer como precio inicial el de mercado, pero si no se lleva a cabo al inicio, se recomendaría hacerlo conforme pase el tiempo –cuando el vendedor manifiesta un claro interés de vender por parte, mostrándose abierto a negociar y vender más barato, como comentaban Quan y Quigley (1991)– pues, solo con mayores variaciones sufridas, en número y, especialmente en rebaja de precio, como aparecía en el estudio de Khezr (2015), es como disminuirá el tiempo de venta.

Y como último componente del modelo, la variable que resulta ser más importante: el interés del comprador. Un inmueble tiene probabilidades de ser vendido siempre que alguien demuestre interés por comprarlo, lo que puede verse reflejado con las apariciones y visitas que recibe y los contactos con el vendedor acerca del mismo. Esta última variable podría completarse integrando también la motivación de compra/venta, imposible en este caso al no disponer de tal información: el vendedor no comenta su motivación por vender y del comprador no se llega a conocer ningún dato. En cualquier caso, es importante comentar que no solo se venden inmuebles que reciben contactos previamente: de la muestra que se dispone, alrededor de un 20% finaliza la transacción a pesar de no tener contacto (28% de chalés y 19% de pisos). Por ello, se intuye entonces que estos no se han producido mediante correo electrónico –vía idealista– sino por fuera: directamente por teléfono o a través de otra fuente.

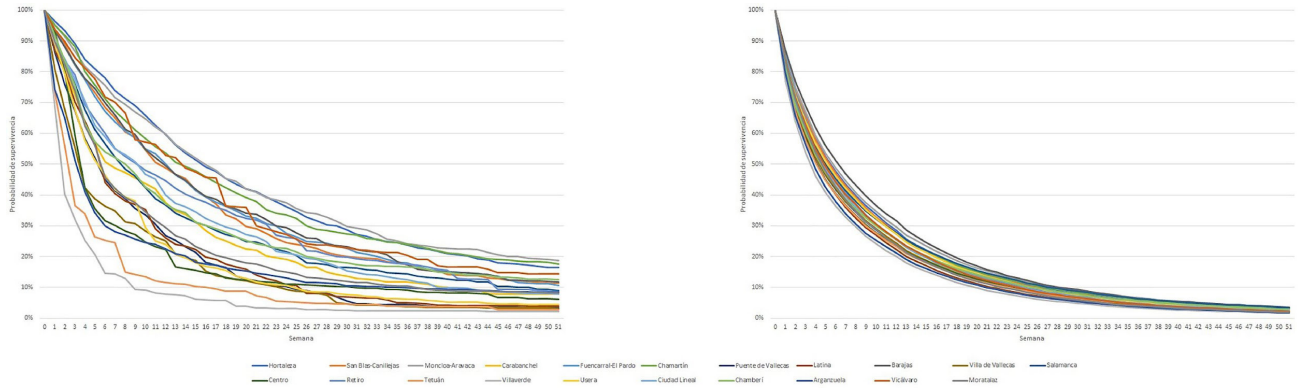
Se muestra un cuadro resumen en la Tabla 8 de los principales hallazgos discutidos en los párrafos anteriores, confirmando en la última columna si se ratifican o refutan las afirmaciones previas de la literatura.

Es importante comentar también que las diferencias de las curvas de supervivencia observadas en la Figura 4 del apartado anterior, podrían tener su explicación no solo en el bajo volumen de chalés en relación al de pisos que se dispone (de ahí que el gráfico de los primeros sea mucho más escalonado) sino en la diferencia dineraria de ambas tipologías: la diferencia de importe de los chalés se sitúa en tres millones de euros y 41 veces según la zona, siendo Villaverde el distrito más barato y Hortaleza, Chamartín y Moncloa-Aravaca de los más caros, vs. 1,2 millones de euros y 15 veces en pisos. La economía juega un papel fundamental en las decisiones de compra, y se ve así reflejado en la Figura 5: los inmuebles que se venden antes son los situados en los barrios del Sur de Madrid, los cuales son, generalmente, las zonas de menor poder adquisitivo.

El aporte de este trabajo con relación a otros previos se sustenta en dos aspectos. En primer lugar, aquellos del ámbito cualitativo, siendo el principal tema que destacar la reponderación de la muestra para adaptarla a la población real estudiada, fundamental para poder aplicar el modelo y no existir sesgos en los

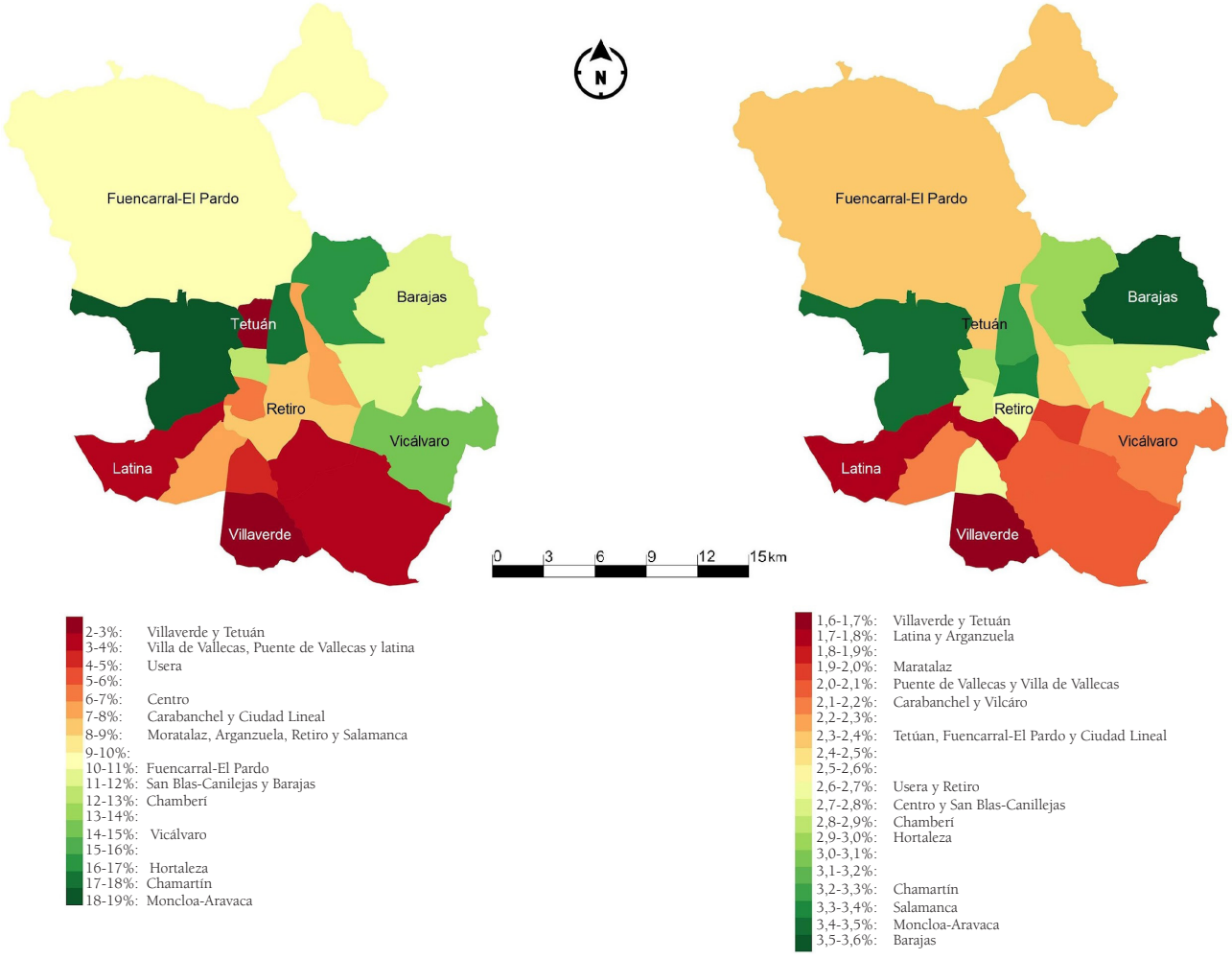
resultados obtenidos: se ha llegado hasta el máximo nivel de detalle posible, dando pesos a cada inmueble registrado como venta del portal, en función de los datos oficiales, según su superficie, barrio y año de venta. También la amplitud de variables utilizadas ha sido mayor, donde hasta 39 variables se han incorporado al modelo final (como se puede observar en la Tabla 6), cubriendo prácticamente la totalidad de los aspectos estudiados previamente en la literatura. Desde el punto de vista cuantitativo, y comparado con el de Cox de Bhuiyan y Hasan (2016), el modelo propuesto obtiene un *C-index* algo inferior (66% vs. 69%, pero hay que remarcar de nuevo la falta de reponderación en su caso), aunque, si comparamos con nuestros mejores, éstos muestran una superioridad notable (76% en chalés y 72% en pisos). El mayor volumen de inmuebles juega un papel importante en estos algoritmos: en nuestro caso se han recopilado decenas de miles de ellos (casi 24 veces más).

Figura 4.
Probabilidad de no venta por distritos para chalés (a la izquierda) y pisos (derecha) sobre la muestra de test.



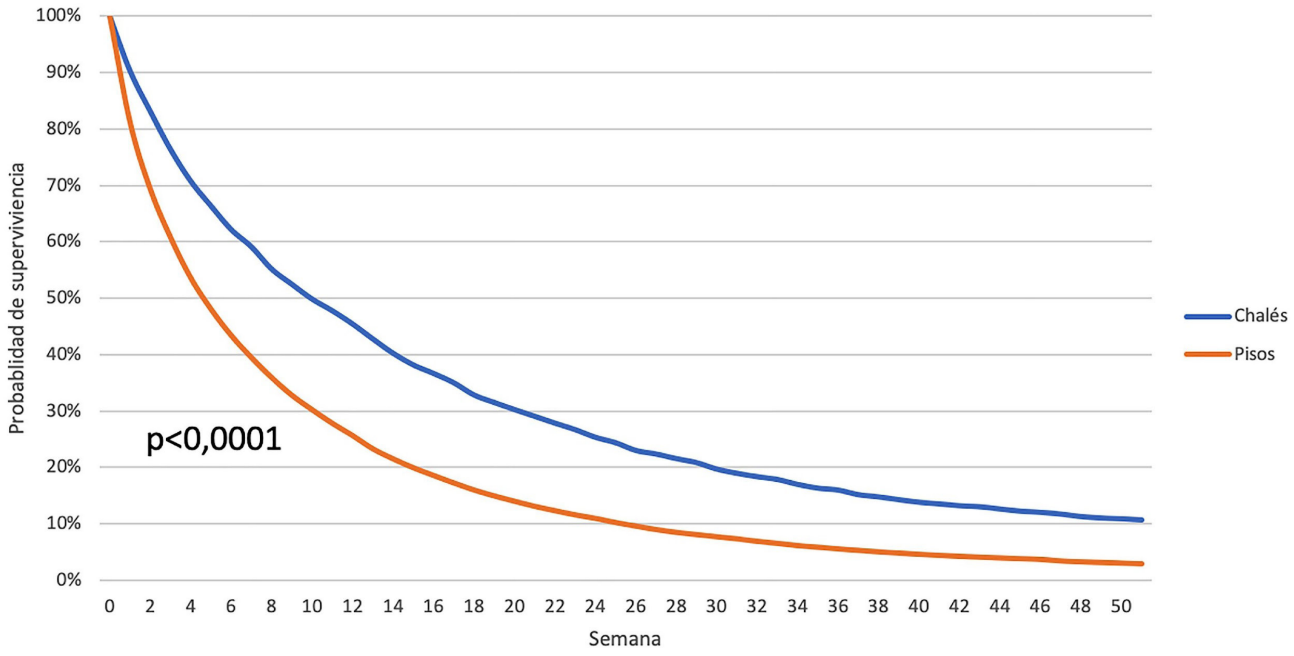
Fuente: elaboración propia.

Figura 5.
Mapa de calor de la probabilidad de no venta en un año para la muestra de test, para chalés (a la izquierda) y pisos (derecha).



Fuente: elaboración propia.

Figura 6.
Curvas de supervivencia (probabilidad de no venta) de Kaplan-Meier y test logrank para chalés y pisos.



Fuente: elaboración propia.

Tabla 8.
Contraste de las variables de la literatura relacionadas con el tiempo de venta.

Autores	Variable	TOM	Este estudio
McCall y Lippman (1984) / Lazear (1986)	Mayor precio de venta	Mayor / Menor en épocas de crisis	Ratificado mayor (el análisis no es en crisis)
Forgey *et al.* (1996)	Propiedades nuevas o de mayor tamaño	Mayor	Ratificado
Anglin *et al.* (2001)	Presencia de piscina o garaje	Menor	Refutado: mayor TOM
Anglin *et al.* (2001) / Bich *et al.* (2020)	DOP	Mayor / Menor	Ratificado mayor
Khezr (2015)	Mayor rebaja de precio en las revisiones de éste	Menor	Ratificado
Cirman *et al.* (2015)	Localización central (diferencia por zonas)	Mayor	Ratificado

Fuente: elaboración propia.

Conclusiones

La aplicación de las técnicas de *machine learning* permite establecer mecanismos de naturaleza económica para establecer índices o referencias que ayuden a corregir las ineficiencias que se producen en el precio, e, incluso, adelantarse a ellas; pero cuya derivada tendrá también un componente social, en la medida que la vivienda es un elemento básico para cualquier ciudadano, en el que además invierte una parte significativa de su patrimonio monetario.

La metodología desarrollada finaliza con un modelo que aúna todos los elementos anteriores, pues emplea algoritmos más innovadores no incorporados hasta ahora en los análisis de supervivencia del mercado de la vivienda, considerando no solo un volumen elevado de datos para un tratamiento masivo mediante técnicas avanzadas, sino soportando la veracidad de los resultados en base a la estadística pública por medio de técnicas estadísticas y econométricas, eliminando así los sesgos que se producen cuando los datos de plataformas de mercado son directamente usados. Todo ello permitirá la elaboración de análisis más razonables y fundamentados en los procesos de compraventa de inmuebles, advirtiendo diferencias tanto físicas como espaciales, que ayudarán a una toma de decisiones óptima, tanto por parte del comprador como del vendedor.

Para mejorar la eficiencia de las transacciones, los participantes en el proceso de compraventa deberían tener el conocimiento de cuáles son los elementos que determinan la evolución de venta de las viviendas y en qué medida lo hacen, cuestión que el presente estudio anima a conocer. Así, permitirán tomar mejores decisiones a los agentes implicados, estableciendo un patrón de comportamiento para conocer la mejor vivienda de entre todas las posibles. Como se ha podido comprobar, no solo influyen las características intrínsecas de las mismas o su localización, sino si el precio está o no en el mercado, las variaciones sufridas por este a lo largo del tiempo y, sobre todo, la capacidad de determinar el potencial interés por parte de posibles compradores, fácilmente obtenible a partir de las búsquedas o contactos recibidos del inmueble en cuestión. Si la operación nace a través de una plataforma de mercado, estas podrían considerar la incorporación de una nueva funcionalidad para dar este servicio a los usuarios. Así, ellos podrían conocer el tiempo de venta estimado en función del precio de publicación o, incluso, proporcionarle este valor al vendedor si desea venderlo en un espacio máximo de tiempo. Sin duda, todo ello podría liberalizar aún más el mercado de la vivienda.

Por tanto, esta metodología escalable, trasladada a otras localizaciones, puede servir para acercar los precios de compra, tanto en precio como en tiempo, lo cual ayudaría a evitar desajustes entre oferta y demanda, así como las numerosas burbujas que en el mercado de la vivienda se han producido: se podrá saber si un determinado inmueble, en una localización concreta, bajo unas características que lo definen, se puede vender o no a un precio definido. Esto favorecerá el acercamiento entre el precio exigido por el vendedor y el ofertado por el comprador, permitiendo también a los posibles agentes encargados de la transacción una mayor eficiencia. Por ello, una herramienta de este tipo, como la comentada previamente, podría contribuir a reducir la brecha en ciertos grupos con menor poder adquisitivo.

Además, la discusión teórica que puede derivarse del documento podría tener relevancia en el contexto de las políticas públicas de asequibilidad. Puede servir para tomar decisiones o hacer leyes, como base o criterio de decisión para poder encauzar las políticas administrativas enfocadas al diseño o la prevención del tema de viviendas, o también en el ámbito fiscal. Quizá en un futuro también ayude a diversos organismos a revisar el proceso de compraventa, dado que podría ayudar en la comprensión de las decisiones de consumo y, tal vez, a regular este mercado tan complejo basándose en este soporte.

Cabe recalcar que, para este estudio, principalmente se han utilizado las características intrínsecas de las viviendas. En posteriores investigaciones, y con un periodo más amplio de estudio, el trabajo se podría completar incorporando cuestiones demográficas o de coyuntura económica, pues se espera que también tengan un papel importante y puedan explicar y ahondar más en los fenómenos aquí recogidos.

Para finalizar, y una vez demostrada la viabilidad de la metodología en un mercado grande y líquido como la capital de España, Madrid, convendría extender la misma a otras regiones españolas para testar si existen variables que condicionen decisiones particulares asociadas a aspectos locales, y así poder identificarlos. Del mismo modo, analizar el comportamiento del alquiler podría ayudar a comprender en mayor medida el mercado de la vivienda, en vista de que los matices propios de él puedan ser diferentes. Por otra parte, se anima a utilizar este tipo de modelos en otros mercados, puesto que podría ser una valiosa fuente de resultados e información tanto para el sector privado como para la formulación de políticas públicas.

Declaración de autoría

David Sánchez Cabrera: conceptualización, curación de datos, análisis formal, metodología, redacción – borrador original.

Julio González Arias: conceptualización, curación de datos, metodología, administración del proyecto, supervisión, redacción – borrador original.

David Rey Blanco: conceptualización, curación de datos, metodología, supervisión, validación, redacción – revisión y edición.

Referencias bibliográficas

Ahmed, A., Ardila, D., Sanadgol, D., y Sornette, D. (2016). Comparing ask and transaction prices in the Swiss housing market. *Swiss Finance Institute Research Paper Series,* (16-80). https://doi.org/10.2139/ssrn.2894404

An, Z., Cheng, P., Lin, Z., y Liu, Y. (2013). How do market conditions impact price-TOM relationship? Evidence from real estate owned (REO) sales. *Journal of Housing Economics, 22*(3), 250-263. https://doi.org/10.1016/j.jhe.2013.07.003

Anglin, P. M., Rutherford, R., y Springer, T. M. (2001). The trade-off between the selling price of residential properties and time-on-the-market: The impact of price setting. *The Journal of Real Estate Finance and Economics, 26,* 95–111. https://doi.org/10.1023/A:1021526332732

Ayuntamiento de Madrid. (2020). *Compra-venta de viviendas*. https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Edificacion-y-vivienda/Mercado-de-la-vivienda/Compra-venta-de-viviendas/?vgnextfmt=default&vgnextoid=9b8db9602f841510VgnVCM1000000b205a0aRCRD&vgnextchannel=22613c7ea422a210VgnVCM1000000b205a0aRCRD

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., y Afonso, C. (2018). Identifying real estate opportunities using machine learning. *8*(11), 2321. https://doi.org/10.3390/app8112321

Bhuiyan, M. y Hasan, M. (2016). *Waiting to be sold: Prediction of time-dependent house selling probability.* 2016 IEEE 3rd International Conference on Data Science and Advanced Analytics (DSAA), 468-477. https://doi.org/10.1109/DSAA.2016.58

Bich, H. N., Trong, H. N., y Thanh, H. T. (2020). The role of listing price strategies on the probability of selling a house: Evidence from Vietnam. *Real Estate Management and Valuation, 28*(2), 63-75. https://doi.org/10.1515/remav-2020-0016

Chen, Y. y Rosenthal, R. W. (1996). On the use of ceiling-price commitments by monopolists. *The RAND Journal of Economics, 27*(2), 207-220. https://www.jstor.org/stable/2555923

Choy, L. H. T. y Ho, W. K. O. (2023). The use of machine learning in real estate research. *Land, 12*(4), 740. https://doi.org/10.3390/land12040740

Cirman, A., Pahor, M., y Verbic, M. (2015). Determinants of time on the market in a thin real estate market. *Engineering Economics, 26*(1). https://doi.org/10.5755/j01.ee.26.1.3905

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Deville, J.-C., Särndal, C.-E., y Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association, 88*(423), 1013-1020. https://doi.org/10.1080/01621459.1993.10476369

Eurostat. (2019). *Distribution of population by tenure status, type of household and income group - EU-SILC survey* [datos]. https://ec.europa.eu/eurostat/databrowser/view/ilc_lvho02/default/table?lang=en

Forgey, F. A., Rutherford, R. C., y Springer, T. M. (1996). Search and liquidity in single–family housing. *Real Estate Economics, 24*(3), 273-292. https://doi.org/10.1111/1540-6229.00691

Geltner, D., Kluger, B. D., y Miller, N. G. (1991). Optimal price and selling effort from the perspectives of the broker and seller. *Real Estate Economics, 19*(1), 1-24. https://doi.org/10.1111/1540-6229.00537

Genesove, D. y Mayer, C. (1997). Equity and time to sale in the real estate market. *The American Economic Review, 87*(3), 255-269.

Glower, M., Haurin, D. R., y Hendershott, P. H. (1998). Selling time and selling price: the impact of seller motivation. *Real Estate Economics, 26*(4), 719-740. https://doi.org/10.1111/1540-6229.00763

Ministerio de Transportes y Movilidad Sostenible. (2023). *Transacciones inmobiliarias (compraventa) [datos].* https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=34000000

Han, L. y Strange, W. C. (2016). What is the role of the asking price for a house? *Journal of Urban Economics, 93,* 115-130. https://doi.org/10.1016/j.jue.2016.03.008

Harrell, F., Califf, R., Pryor, D., Lee, K., y Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA, 247*(18), 2543-2546. https://doi.org/10.1001/jama.1982.03320430047030

Haurin, D. (1988). The duration of marketing time of residential housing. *Real Estate Economics, 16*(4), 396-410. https://doi.org/10.1111/1540-6229.00463

Instituto Nacional de Estadística. (2002a). Cambios metodológicos EPA-2002. *Cifras INE. Boletín Informativo del Instituto Nacional de Estadísticas,* (3). https://www.ine.es/epa02/cifrasine_epa02.pdf

Instituto Nacional de Estadística. (2002b). *Método de reponderación aplicado en la EPA.* https://www.ine.es/epa02/documento_tecnico.pdf

Instituto Nacional de Estadística. (2019). *Hogares por régimen de tenencia de la vivienda y tipo de hogar.* Autor. https://www.ine.es/jaxiT3/Tabla.htm?t=9996&L=0

Instituto Nacional de Estadística. (2023a). *Estadística de transmisiones de derechos de la propiedad. Compraventa de viviendas según régimen y estado.* Autor. https://www.ine.es/jaxiT3/Tabla.htm?t=6150&L=0

Instituto Nacional de Estadística. (2023b). *Series históricas de población desde 1996. Cifras oficiales de la Revisión anual del Padrón municipal a 1 de enero de cada año.* Autor. https://www.ine.es/jaxiT3/Tabla.htm?t=29005&L=0

International Valuation Standards Council. (2022). *International valuation standards.* Autor. https://www.rics.org/content/dam/ricsglobal/documents/standards/ivsc_effective_31_jan_2022.pdf

Johnson, K., Benefield, J., y Wiley, J. (2007). The probability of sale for residential real estate. *Journal of Housing Research, 16*(2), 131-142. https://doi.org/10.1080/10835547.2007.12091978

Kang, H. B. y Gardner, M. J. (1989). Selling price and marketing time in the residential real estate market. *Journal of Real Estate Research, 4*(1), 21-35. https://doi.org/10.1080/10835547.1989.12090570

Kaplan, E. L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481. https://doi.org/10.1080/01621459.1958.10501452

Khezr, P. (2015). Time on the market and price change: the case of Sydney housing market. *Applied Economics, 47*(5), 485-498. https://doi.org/10.1080/00036846.2014.972549

Kleinbaum, D. y Klein, M. (1996). *Survival analysis: a self-learning text.* Springer.

Knight, J. R. (2002). Listing price, time on market, and ultimate selling price: Causes and effects of listing price changes. *Real Estate Economics, 30*(2), 213-237. https://doi.org/10.1111/1540-6229.00038

Lazear, E. (1986). Retail pricing and clearance sales. *The American Economic Review, 76*(1), 14–32.

Li, W.-F. (2015). The impact of pricing on time-on-market in high-rise multiple-unit residential developments. *Pacific Rim Property Research Journal, 10*(3), 305-327. https://doi.org/10.1080/14445921.2004.11104165

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports, 50*(3), 163-70.

McCall, J. J. y Lippman, S. A. (1984). *An operational measure of liquidity*. University of Konstanz, Department of Economics.

Ngai, L. y Tenreyro, S. (2014). Hot and cold seasons in the housing market. *The American Economic Review, 104*(12), 3991–4026.

Oxford Spring School. (2007). *An introduction to event history analysis*. https://spia.uga.edu/faculty_pages/rbakker/pols8501/OxfordTwoNotes.pdf

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research, 21*(212), 1–6.

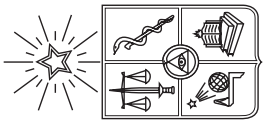
Quan, D. C. y Quigley, J. M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics, 4*(2). https://doi.org/10.1007/bf00173120

Scofield, D. y Devaney, S. (2017). What sells in a crisis? Determinants of sale probability over a cycle and through a crash. *Journal of Property Investment & Finance, 35*(6), 619-637. https://doi.org/10.1108/jpif-02-2017-0013

Selcuk, C. (2012). *Motivated sellers predatory buyers*. Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/36226/

Vergara-Perucich, J. (2021). Precios y financierización: evidencia empírica en mercado de la vivienda del Gran Santiago. *Revista INVI, 36*(103), 137–166. https://doi.org/10.4067/S0718-83582021000300137

revista INVI



Revista INVI es una publicación periódica, editada por el Instituto de la Vivienda de la Facultad de Arquitectura y Urbanismo de la Universidad de Chile, creada en 1986 con el nombre de Boletín INVI. Es una revista académica con cobertura internacional que difunde los avances en el conocimiento sobre la vivienda, el hábitat residencial, los modos de vida y los estudios territoriales. Revista INVI publica contribuciones originales en español, inglés y portugués, privilegiando aquellas que proponen enfoques inter y multidisciplinares y que son resultado de investigaciones con financiamiento y patrocinio institucional. Se busca, con ello, contribuir al desarrollo del conocimiento científico sobre la vivienda, el hábitat y el territorio y aportar al debate público con publicaciones del más alto nivel académico.

Sitio web: http://www.revistainvi.uchile.cl/
Correo electrónico: revistainvi@uchilefau.cl