BCU

Predicción de precios de vivienda

Aprendizaje estadístico con datos de oferta y transacciones para la ciudad de Montevideo

Pablo Picardo

Nº 002 - 2019

Documento de trabajo
ISSN 1688-7565

Predicción de precios de vivienda
Aprendizaje estadístico con datos de oferta y transacciones para la ciudad de Montevideo[☆]

Pablo Picardo [a 1]

a *Banco Central del Uruguay (Inveco), 777 Diagonal J.P. Fabini 11100 Montevideo, Uruguay*

Documento de trabajo del Banco Central del Uruguay 002-2019

Autorizado por: Jorge Ponce

Resumen

En este trabajo se presentan modelos predictivos para el precio de un activo de difícil valuación como la vivienda. Se utilizan dos fuentes de datos para la ciudad de Montevideo: una proveniente de sitios web (a través de web *scraping*) y otra de registros administrativos de transacciones. Se implementan tres modelos fácilmente replicables: modelo lineal, árbol de regresión y bosques aleatorios. Los resultados arrojan una mejor performance del modelo de bosques aleatorios respecto al modelo lineal hedónico, ampliamente difundido en la literatura. Se busca incorporar al análisis de predicción de precios una metodología aún escasamente difundida a nivel nacional, implementada en el software R y poner a disposición una nueva base de datos.

JEL: C10, C18, C52, C81, R31
Palabras clave: precios de vivienda, aprendizaje estadístico, bosques aleatorios, CART, valuación de activos, precios online, datos geo-referenciados

[☆]Agradezco a la Dirección General de Registros por permitirme acceder a los datos administrativos de transacciones y muy especialmente a Bruno García, quien me ayudó a recolectar los datos de la web y a Joselina Davyt, quien me brindó una excepcional ayuda con temas vinculados al software R. Agradezco a todos mis colegas de Investigaciones Económicas del Banco Central del Uruguay, que me han apoyado en todo lo referente a este trabajo. Por último agradezco muy especialmente a Natalia da Silva y Fernando Borraz, quienes me apoyaron en el desarrollo de la idea y en sus vaivenes. Los errores son de mi exclusiva responsabilidad y no comprometen a ninguna de las instituciones en las que trabajo.

[1] Correo electrónico: *ppicardo@bcu.gub.uy*

Índice

1. Introducción

El presente trabajo aporta una nueva base de datos de vivienda para la ciudad de Montevideo e implementa un enfoque de aprendizaje estadístico al desafío de predicción de precios de inmuebles. En primer lugar, se recopilan datos de un sitio web y de registros administrativos. En segundo lugar se modela el precio del inmueble con técnicas de aprendizaje estadístico, de creciente aplicación en la literatura estadística, computacional e incipientemente económica. Esto introducirá la discusión de la complementariedad entre modelos de estimación y modelos de predicción.

El mercado inmobiliario tiene un rol clave en la actividad económica, como indica Leamer (2007): *Housing is the business cycle*. Si bien esto no es algo reciente, este mercado ha jugado roles centrales en las crisis económicas y financieras (Mooya, 2016); la última crisis internacional de 2007/08 es un ejemplo. Posteriormente, en los países latinoamericanos, el gran flujo de capitales afectó los precios de los activos internos, lo que generó preocupación por los eventuales desvíos de fundamentos, como bien documentan Licandro y Ponce (2015). Uno de los mayores tópicos referidos al mercado inmobiliario se refiere a la valuación de los bienes y la detección de sobreprecios (Zhu, 2014, Mooya, 2016, Fischer, 2017) ya sea para los interesados en comprar y vender como para las instituciones financieras que mantienen en su activo derechos sobre propiedades inmobiliarias y su solvencia depende del valor de las mismas. Esto último explica parte del interés de los organismos reguladores de las entidades financieras en este mercado.

La característica distintiva del mercado inmobiliario es que la vivienda tiene la doble función de ser un bien de inversión y una fuente de utilidad para quienes lo usan. La vivienda es un *bien preferente* (Fischer, 2017) y su precio es importante a la hora de definir cómo hacer que las personas accedan a ese bien. Adicionalmente, la vivienda tiene una extensa vida útil y un relativamente largo proceso de producción, lo que genera rigideces en la oferta. Esto deriva en que el precio tenga un comportamiento diferente al de cualquier otro bien, lo que es exacerbado por el contexto uruguayo de dolarización (Drenik y Pérez, 2017). En cualquier caso, la valuación de este bien heterogéneo es un desafío relevante (Mooya, 2016).

En Uruguay, como en muchos países, la vivienda es el activo principal de los hogares (Encuesta Continua de Hogares). Además, debido a la poca profundidad del mercado de capitales, la compra de vivienda es una de las escasas opciones de inversión local. Se trata también de un activo/riqueza cuyo precio incide directamente en una gran parte de la población. A modo de ejemplo, según la Encuesta Continua de Hogares (2014) el 59 % de los hogares son propietarios de su vivienda, ya sea habiendo pagado la misma en su totalidad (50 %) o no (9 %). Por otra parte, este bien se encuentra distribuido más equitativamente que otros (De Rosa et al., 2016). En efecto, el precio de la vivienda es importante a la hora de la valuación de la riqueza de un número importante de personas y/o instituciones (propietarios, aspirantes a propietarios, inversores, instituciones financieras, organismos gubernamentales, etc.).

En línea con lo anterior y como indican Ponce y Tubio (2013), el financiamiento de la vivienda forma parte del principal pasivo de los hogares y, aunque el mercado hipotecario es aún poco profundo en Uruguay (representa aproximadamente 15 % del total de crédito), es un canal directo hacia la estabilidad financiera (Landaberry y Tubio, 2015). El banco predominante, en este mercado y en Uruguay, es el Banco Hipotecario del Uruguay (BHU), sin embargo, en los últimos años los bancos privados han aumentando su participación en el mercado y han surgido más jugadores (fideicomisos financieros e inversores extranjeros) así como nuevas regulaciones: Ley de Vivienda Promovida [1].

En otro orden y más allá de la vivienda, los avances tecnológicos permiten acceder a mayor cantidad de datos, de mayor calidad y con mayor rapidez. Esto genera oportunidades de trabajo transversales entre la computación, la estadística y la economía (Varian, 2014). En lo que refiere al mercado inmobiliario, se destaca una tendencia creciente a la centralización y digitalización de datos en bases únicas y la posibilidad de geo-referenciación. Este trabajo pretende, de forma simple, aprovechar esta ventaja.

Las secciones que siguen se estructuran de la siguiente manera: en primer lugar se hace un breve repaso de los antecedentes sobre trabajos de predicción en el mercado inmobiliario y de los que utilizan de técnicas de aprendizaje estadístico aplicadas al mismo. En segundo lugar se define un marco teórico y la estrategia metodológica. En la sección 4 se presentan los datos y luego se exponen los resultados. Por último, se realizan comentarios finales y se indican líneas de trabajo a futuro.

2. Antecedentes

Este trabajo consiste en la aplicación de modelos de aprendizaje estadísticos en el mercado de vivienda de la ciudad de Montevideo. En torno a este tema, los antecedentes se pueden dividir entre los trabajos que tratan sobre el mercado de vivienda en Uruguay y los que aplican técnicas de aprendizaje estadístico para la predicción de precios de vivienda. Para este último punto se encontró un solo trabajo nacional y numerosas referencias internacionales.

Dentro de la segunda categoría, se destaca el trabajo de Mullainathan y Spiess (2017), que realiza una introducción a los modelos de aprendizaje estadístico mediante un ejemplo con datos de precios de vivienda. En éste, ilustran diferentes modelos de predicción y muestran su performance, destacando la mejor performance predictiva de los modelos de aprendizaje. Este artículo, al igual que Athey (2018) y Varian (2014) fueron la principal motivación para incorporar técnicas de aprendizaje estadístico y resultan referencias básicas en la utilización de modelos de aprendizaje estadístico en Economía.

En relación a literatura sobre el mercado inmobiliario a nivel local, en Ponce y Tubio (2013) se propone una aplicación empírica del modelo lineal de precios hedónicos con

[1] Ver: González-Pampillón (2017), Berrutti Rampa (2016), García López (2018).

datos del sitio web buscandocasas.com.uy. A su vez, Landaberry y Tubio (2015) realizan una estimación del modelo hedónico con datos de transacciones y características de catastro. Trabajos como los mencionados, se basan en la aplicación del modelo que más abunda en Economía (puede ser profundizado en la revisión bibliográfica de Herath y Maier (2010)).

Los trabajos de Carlomagno y Fernández (2007), Domínguez et al. (2016), Lanzilotta y Veneri (2016) son antecedentes relevantes de la bibliografía nacional que se enfocan en los determinantes de los precios de la vivienda. En el caso de los dos primeros desde un punto de vista macroeconómico y en el tercero desde una perspectiva microeconómica y hedónica. Se destaca el interés de estimar el *efecto barrio* con herramientas de la econometría espacial: se pone el foco en el peso de la ubicación de los inmuebles en el precio, al igual que en Kiel y Zabel (2004), que enfatiza el rol de la ubicación y las características asociadas a ella. En línea con la utilización de la econometría espacial, el trabajo de González-Pampillón (2017) utiliza la ubicación exacta de los inmuebles con los datos de transacciones de la DGR y así estudia el efecto de la Ley de Vivienda Promovida en el precio de los inmuebles.

La referencia nacional que implementa técnicas de aprendizaje es el trabajo de Goyeneche et al. (2017), en el que se utiliza una base de datos de tasaciones del BHU con el objetivo de predecir el precio contado de un inmueble, definido como el asignado por un tasador. Estos autores utilizan diversas metodologías, considerando la dimensión temporal, que luego agregan mediante el método *Stacking* (Breiman, 1996). La modelización, según los autores, *no sustituye el trabajo de un tasador que se asocia más a un arte que muchas veces depende de aspectos no cuantificables y/o no observables*. En este trabajo la utilización del modelo de Stacking reduce a la mitad la Raíz del Error Cuadrático Medio (RECM o RMSE en inglés) respecto al modelo Semiparamétrico Espacial Dinámico (modelo lineal hedónico con efectos espaciales). Lo anterior implica que el error se reduce de USD 24.000 a USD 12.000.

Por otra parte, en el reciente artículo de Čeh et al. (2018) se realiza un trabajo similar al presentado en esta tesis. En éste comparan la performance predictiva de un modelo de Bosques Aleatorios en relación a una regresión lineal hedónica para el precio de los apartamentos en Liubliana, Eslovenia. Estos autores utilizan diversas variables geográficas (distancias y proximidades a lugares “deseables” e “indeseables”), una tendencia creciente en la literatura sobre precios de los inmuebles. En el trabajo para Liubliana, el modelo lineal hedónico produce un Error Porcentual Absoluto Medio (EPAM o MAPE en inglés) del orden de 17 % en una muestra de testeo, mientras que el modelo de Bosques Aleatorios alcanza un MAPE del orden del 7 %. Esto significa que este método reduce a más de la mitad el error.

En general, los antecedentes nacionales sobre predicción y estimación de determinantes de precios de la vivienda se encuentran con una barrera común: los datos. El trabajo de Ponce y Tubio (2013) cuenta con una base de transacciones inmobiliarias de la Dirección General Impositiva sin ubicación exacta y con pocas características (las de

catastro). Este trabajo cuenta con datos de ofertas también, aunque de un período restringido de tiempo y con relativamente pocas observaciones. En el caso de los trabajos de Lanzilotta y Veneri (2016), Domínguez et al. (2016) y Carlomagno y Fernández (2007), se basan en datos agregados del Instituto Nacional de Estadística. En Goyeneche et al. (2017) se utilizan datos del BHU con gran cantidad de características de los inmuebles y confiabilidad, sin embargo refieren a un segmento específico del mercado.

La mayoría de artículos de aprendizaje estadístico aplicado a precios de inmuebles se asocia a disciplinas como la computación, la geografía y la estadística. Además de Goyeneche et al. (2017), Čeh et al. (2018), pueden encontrarse aplicaciones al mercado inmobiliario en: Fan et al. (2006) que utiliza árboles de decisiones, Wang et al. (2014) que aplica Máquinas de Vectores de Soporte (Support Vector Machines), Chiarazzo et al. (2014) y Selim (2009) que proponen modelos de Redes Neuronales; el primero de ellos con variables ambientales y de ubicación. Además, en Park y Bae (2015) se utilizan métodos menos difundidos.

Los diversos modelos de aprendizaje estadístico se pueden encontrar resumidos en dos referencias básicas: Friedman et al. (2001) y James et al. (2013). Uno de los modelos que más aparece en la literatura y que se utilizarán aquí son los Árboles de Regresión y Clasificación (CART por sus siglas en inglés) y su agregación en Bosques Aleatorios (*Random Forest*).

Por último, existen aplicaciones no académicas en plataformas como kaggle.com donde empresas ponen a disposición bases datos y problemas reales de predicción a resolver y los usuarios compiten por el premio ofrecido al que mejor lo resuelva. Se destaca el desafío de Zillow.com, una página norteamericana de venta y alquiler de propiedades que ofrece un algoritmo de valuación. A su vez, existen en el sector privado uruguayo productos de este estilo aunque los detalles técnicos y el desempeño de los modelos no se divulgan.

3. Marco teórico y metodología

En la siguiente sección se presenta una breve descripción del marco teórico sobre precios en el mercado de vivienda a partir de un modelo lineal hedónico y con un enfoque de la Teoría del Aprendizaje Estadístico aplicado al problema de predicción del precio. A su vez, se presenta la estrategia metodológica.

Se desarrolla brevemente cada uno de los modelos y la estrategia para realizar la validación cruzada y la comparación entre ellos. Lo anterior incluye las decisiones respecto a los datos y variables que se utilizan, los parámetros de los modelos y las medidas de performance.

3.1. ***Modelo lineal hedónico***

El modelo de precios hedónico fue originalmente introducido por Griliches (1961) para el precio de los automóviles y luego fue desarrollado teórica y conceptualmente por Rosen (1974). Los modelos de precios hedónicos plantean que el precio de una bien puede ser determinado en función de sus características observables. Éstas tienen un precio implícito, el cual puede ser aproximado mediante los coeficientes de un modelo de regresión. En el caso de las viviendas, se trata de bienes heterogéneos compuestos por diferentes características cuya valoración marginal puede ser derivada mediante estos modelos (Rosen, 1974). A las características propias de la vivienda pueden agregarse factores asociados al equilibrio del mercado (stock disponible, por ejemplo), características sociodemográficas de la ubicación (edad, ingreso, etc.) y atributos geográficos, como se menciona en De Bruyne y Van Hove (2013). En el ámbito nacional Lanzilotta y Veneri (2016), Ponce y Tubio (2013) y Landaberry y Tubio (2015) son algunos ejemplos en los que se desarrollan este tipo de modelos.

Así, es posible considerar un modelo de regresión lineal para predecir el precio de un inmueble en base a sus características (versión más simple) de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i \tag{1}$$

donde Y_i es el precio de la vivienda i, las $X_1 ... X_p$ refieren a los p atributos de la vivienda i, los $\beta_1 ... \beta_p$ son los coeficientes a estimar y ϵ_i el término de error, independiente e idénticamente distribuido (iid) $N(0, \bar{\sigma})$. La estimación se realiza mediante Mínimos Cuadrados Ordinarios[2].

Estos modelos parten del supuesto de que el precio del inmueble se puede descomponer en una suma del precio de sus características o atributos, se trata de un modelo de regresión lineal donde las covariables son los atributos y la variable de respuesta es el precio del inmueble. Entre las fortalezas del modelo, como se indica en Ponce y Tubio (2013) se encuentra el control por calidad y composición de los bienes, la eficiencia en el uso de la información, la estimación directa de los precios sombra, la posibilidad de construir índices y monitorear desvíos y la facilidad para identificar heterogeneidades. Como desventajas, asociadas al tema de inferencia y estimación, se visualizan problemas de variables omitidas y endogeneidad, sesgos de selección y multicolinealidad.

Es posible tener buenas predicciones a partir de estos modelos lineales, sin embargo, elegir un modelo lineal implica tomar algunas decisiones respecto a la cantidad de variables y sus interacciones además de ciertas restricciones/supuestos. Otros métodos dentro de lo que llamamos aprendizaje estadístico buscan, a partir de todas las variables y sus interacciones, mejorar la predicción respecto del modelo lineal.

[2]No se desarrolla el modelo lineal ni la estimación por mínimos cuadrados por motivos de espacio. Una descripción en el marco del tema de aprendizaje estadístico puede encontrarse en James et al. (2013).

3.2. ***Modelos de aprendizaje estadístico***

El aprendizaje estadístico refiere a un amplio conjunto de herramientas y métodos para comprender datos con un foco especial en la predicción (James et al., 2013). Los métodos de aprendizaje se pueden dividir en *supervisados* y *no supervisados*. En el primer caso tienen como fin predecir una variable objetivo (*output*) en base a *inputs*. En el segundo caso, se tienen *inputs* pero no *output* y se trata de encontrar relaciones y estructura en los datos (como por ejemplo en análisis de *clusters*). En consonancia con Mullainathan y Spiess (2017), el aprendizaje supervisado, que es el que se utilizará en el trabajo y que incluye al modelo lineal, toma una función de pérdida $L(\hat{y}, y)$ como *input* y busca una función $\hat{f}$ que minimice la pérdida esperada $E_{(y,x)}[L(\hat{f}(x), y)]$ en una observación nueva, fuera de la muestra. Esto es, se define un modelo, se *entrena* en una base de datos y luego se *testea* en una nueva, diferente a la que se entrenó.

La clave para utilizar estos modelos está en definir problemas donde la predicción ($\hat{y}$) sea lo más relevante por sobre los efectos fijos o marginales ($\hat{\beta}$). Justamente, el mejor rendimiento y flexibilidad de estos modelos van de la mano con una menor interpretabilidad de sus resultados en términos causales (James et al., 2013) y riesgos de sobreajuste (*overfitting*). Esto constituye una de las principales críticas y riesgos al utilizar alguno de estos modelos (Abadie y Kasy, 2017). No obstante, los desarrollos más recientes han tratado de dar un marco conceptual para la inferencia causal (Athey, 2018).

Usualmente, en los trabajos empíricos en Econometría y Economía, primero se especifica un modelo, se lo estima en toda la base de datos y se confía en la teoría estadística para obtener intervalos de confianza para los parámetros estimados. De esta forma, el foco está en los efectos estimados del modelo más que en el poder predictivo del mismo (Athey, 2018). Por lo tanto, se pueden tomar a los métodos de aprendizaje estadístico, donde el foco está en la predicción a partir de ejercicios empíricos, como complementarios a lo que se denomina Econometría clásica. En este trabajo, se utilizan árboles de regresión y clasificación (CART), un método de aprendizaje estadístico introducido por Breiman et al. (1984). Éste se basa en construir predictores a partir de árboles tanto en regresión (si se trata de modelar una variable continua) como en clasificación (si la variable a predecir es categórica). El principio general de este método es particionar recursivamente el espacio de variables explicativas X_i de forma binaria y así determinar sub-particiones óptimas para la predicción. Este método tiene la virtud de ser de fácil interpretación, comunicación y transparente. Sin embargo, en la mayoría de los casos tiene peor performance predictiva que un modelo lineal y una serie de falencias asociadas a la estabilidad de las predicciones (Genuer y Poggi, 2017). Esto se puede resolver en gran medida con la introducción del método agregativo de bosques aleatorios. Éste es más reciente (Breiman, 2001) y conceptualmente implica una agregación de muchos árboles donde se toma la media de las predicciones individuales (para el caso de regresión) o el voto mayoritario (en el caso de clasificación) como criterio de partición. Este método es un tipo de agregación que forma parte de *Bootstrap Aggregation* o *Bagging* y se desarrollará a continuación en base a tres textos de referencia: James et al. (2013), Friedman

et al. (2001), Genuer y Poggi (2017) (además de Breiman et al. (1984), Breiman (2001)).

3.2.1. CART

Considerando que poseemos predictores $X_1, X_2, ..., X_p$ y una variable respuesta continua (a predecir), un árbol de regresión se construye en dos etapas (James et al., 2013):

1. Se divide el espacio del predictor, esto es, el set de valores posibles para $X_1, X_2, ..., X_p$ en J regiones distintas y no superpuestas $R_1, R_2, ..., R_j$.
2. Para cada observación que cae dentro de la región R_j se realiza el mismo paso anterior con una predicción equivalente a la media de la variable respuesta en R_j.

Supongamos que en la etapa 1 obtenemos dos regiones (R_1 y R_2) y que la media de la respuesta en la región R_1 es 100.000 y en R_2 es 200.000. Entonces, para una observación dada $X = x$, si $x \in R_1$ la predicción será de 100.000, de lo contrario, si $x \in R_2$, la predicción será de 200.000.

La construcción de los subespacios R_j se realiza con particiones binarias de los predictores de forma tal que se minimice el Error Cuadrático Medio (ECM) de la variable a predecir dentro de la muestra en cuestión:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{2}$$

Donde $\hat{y}_{R_j}$ es la media de la variable a predecir en la muestra de entrenamiento en la partición j. Sin embargo, dado que es poco factible considerar todas las particiones posibles, el método opta por la partición recursiva o de arriba hacia abajo (*top-down*). Éste comienza en una única región donde se encuentran todas las observaciones y cuya predicción es simplemente la media de la variable a predecir. Luego, se realiza la primera partición binaria tomando en cuenta el predictor que produzca el menor ECM.

La partición binaria recursiva implica seleccionar el predictor X_j y luego el punto de corte s para realizar la partición en dos regiones: $[X|X_j < s]$ y $[X|X_j \geq s]$ que lleven a la mayor reducción del ECM. En detalle, $\forall \quad j \quad \text{y} \quad s$, se define el par de semiplanos:

$$R_1(j, s) = [X|X_j < s] \quad \text{y} \quad R_2(j, s) = [X|X_j \geq s], \tag{3}$$

y se buscan los valores de j y s que minimizan la siguiente expresión:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{4}$$

Donde $\hat{y}_{R_1}$ es la media de la variable respuesta en $R_1(j, s)$ y $\hat{y}_{R_2}$ la correspondiente a $R_2(j, s)$. El paso siguiente es repetir el proceso, a partir de los dos *nodos* (región R_1 y R_2)

anteriormente generados; ya no el espacio total. Se busca nuevamente el mejor predictor y el mejor punto de corte para particionar los datos y lograr minimizar el ECM a partir de las regiones creadas. Así, el proceso continúa hasta indicar un criterio de parada que puede indicarse como un mínimo de observaciones en cada región.

Una vez que las regiones $R_1, R_2, ..., R_j$ han sido definidas, se realiza la predicción (en una muestra de testeo independiente) considerando la media del valor de la variable de interés en las observaciones de entrenamiento dentro de la región a la que la observación corresponde.

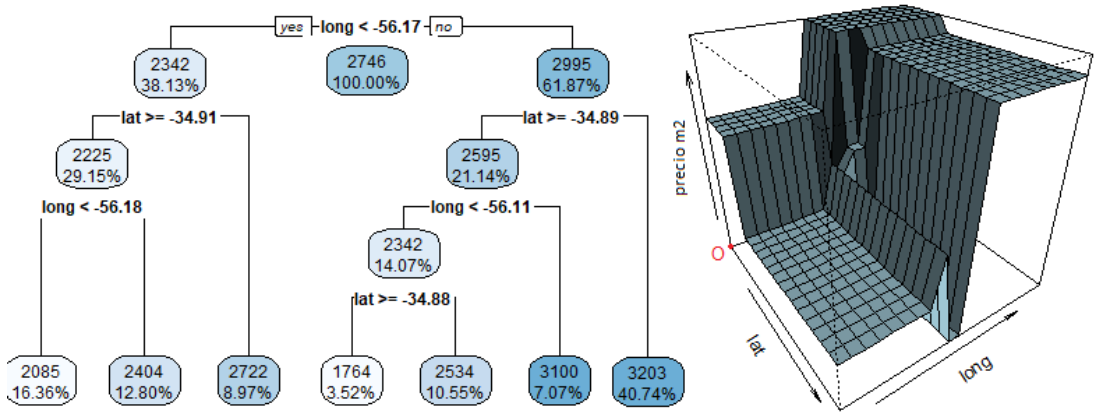
El proceso descrito puede producir excelentes predicciones en la muestra de entrenamiento, sin embargo, es muy probable que produzca sobreajuste y lleve a una pobre performance predictiva en una muestra independiente. En efecto, un árbol con menos particiones puede llevar a menor varianza y mejor interpretación de resultados. Ante esto hay dos alternativas, la primera consta en construir un árbol de un tamaño tal que la disminución del ECM cuando se realiza cada partición exceda algún umbral definido. Esta estrategia dará como resultado árboles más pequeños, pero no considera la posibilidad de que una división aparentemente sin valor al principio del árbol podría ir seguida de una muy buena división, es decir, se podría perder una partición que conduciría a una gran reducción en el ECM más adelante.

La segunda alternativa consiste en construir un árbol *maximal* T_{max} y luego realizar una *poda* hacia atrás para obtener un sub-árbol. Para seleccionar el sub-árbol, se toma en cuenta al parámetro α de costo-complejidad. Para cada valor de α, corresponde un sub-árbol $T \subset T_{max}$ tal que:

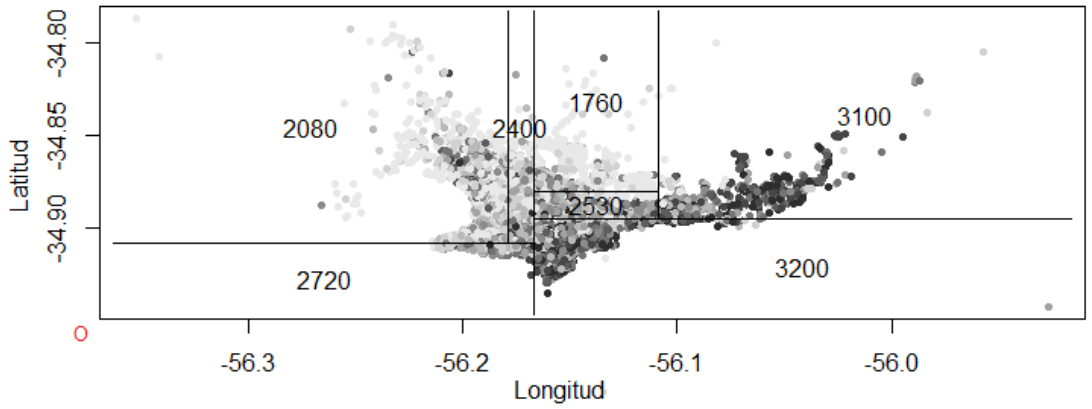
$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{5}$$

es lo menor posible. $|T|$ indica el número de nodos terminales del árbol T, R_m es el espacio formado por los predictores (de forma rectangular) correspondiente al nodo terminal mth, y $\hat{y}_{R_m}$ es la predicción asociada con el nodo R_m, que es la media de las observaciones en R_m (considerando la muestra de entrenamiento). De esta forma, α controla el *trade-off* entre la complejidad del sub-árbol y el ajuste a los datos de entrenamiento. Cuando $\alpha = 0$, el sub-árbol es igual al T_{max}.

A modo de ejemplo, a continuación se presenta un árbol de regresión para predecir el precio del metro cuadrado ofertado de un apartamento para la ciudad de Montevideo. Se toman en cuenta solamente dos predictores: latitud y longitud. Se busca realizar particiones considerando dos variables geográficas y así generar regiones del precio por metro cuadrado. El parámetro α es 0.01, que es el que toma por defecto el paquete RPART (Therneau y Atkinson, 2018). Se opta por modelar la variable precio por m^2 en este ejemplo porque conceptualmente más fácil de interpretar.



Una lectura posible del árbol es la siguiente: el precio promedio de toda la muestra es USD 2.746. La primera partición se realiza a partir de longitud (divide entre oeste y este en este caso) con el valor -56,17. Las propiedades al oeste de la partición (menor longitud que -56,17) tienen un precio promedio de USD 2.342, las ubicadas al este (longitud mayor que -56,17) tienen un precio promedio de USD 2.995. Las particiones siguientes siguen el mismo criterio. Los porcentajes indicados en cada *hoja* refieren a la cantidad de observaciones dentro de la misma. La imagen de la derecha dispone en 3D el árbol. Tomar el punto O como referencia.



Intensidad del color en el gráfico refiere al rango que definen los deciles del precio del metro cuadrado, los puntos más oscuros representan los inmuebles más caros (precio varían desde USD 540 a USD 5.000). Los valores indicados refieren al precio por metro cuadrado predichos por el árbol (se redondean las decenas).

Figura 1: Partición con predictores geográficos a partir de un árbol de regresión para predecir el precio ofertado del metro cuadrado de apartamentos en Montevideo

Fuente: Api de Mercadolibre, elaboración propia

3.2.2. Bosques Aleatorios

Los árboles presentados anteriormente sufren de alta varianza. Esto significa que si se divide una muestra aleatoriamente entre entrenamiento y testeo y luego se ajusta un árbol para ambas muestras, el resultado puede ser muy diferente. En este contexto surgen los métodos agregativos que buscan reducir esta varianza.

El método de bosques aleatorios, denominado en inglés *Random Forest* es un caso particular del método de Agregación por Boostrap (*Boostrap Aggregation o Bagging*). Es necesario definir éste en primer lugar.

Siguiendo el manual de James et al. (2013), recordamos que de un set de n observaciones independientes $Z_1, Z_2, ..., Z_n$ cada una con varianza θ^2, la varianza de la media $\bar{Z}$ está dada por θ^2/n. En otras palabras, promediar un conjunto de observaciones reduce la varianza. De esta forma, se puede calcular $\hat{f}^1(x)$,$\hat{f}^2(x)$,...,$\hat{f}^B(x)$ utilizando B bases de entrenamiento, promediarlos y así obtener un modelo de baja varianza dado por:

$$\hat{f}_{prom}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{6}$$

No obstante, esto no se puede realizar ya que no se cuenta con B muestras de entrenamiento. Por ello, se recurre al *boostrap* tomando muestras aleatorias con reposición a partir de la base de entrenamiento. De esta forma se generan B muestras de entrenamiento diferentes a través del método de *bootstrap* y se obtiene lo siguiente, que se denomina *bagging*:

$$\hat{f}_{bagging}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{7}$$

Si bien lo anterior puede utilizarse para cualquier método, es muy útil en el caso de los árboles. Para aplicar este método con árboles, simplemente construimos B árboles utilizando conjuntos de entrenamiento tomados a partir de *bootstrap*, y promediamos las predicciones resultantes. Los árboles a considerar no se podan, por lo tanto, cada árbol individual tiene una alta varianza, pero un bajo sesgo. Se ha demostrado que este método proporciona grandes mejoras en la precisión al combinar cientos o incluso miles de árboles en un solo procedimiento (la cantidad estándar es 500) (James et al., 2013).

Los Bosques Aleatorios introducen un cambio en el método anterior. Al igual que en *Bagging*, construimos un bosque de árboles de decisión en las muestras de entrenamiento tomadas con *Bootstrap*. Sin embargo, cuando se construyen estos árboles, cada vez que se considera una división, se elige una muestra aleatoria de m predictores como candidatos tomados del conjunto completo de p predictores. El valor estándar de m es $\sqrt{p}$ y así lo considera el paquete *randomForest* (Liaw y Wiener, 2002). Observar que si se elige $m = p$, estamos ante el método *Bagging*.

La principal característica del método de Bosque Aleatorio respecto del *Bagging* es que el Bosque Aleatorio no considera la mayoría de los predictores disponibles en cada división del árbol. El razonamiento que justifica este mecanismo es el siguiente: supongamos que hay un predictor muy fuerte en el conjunto de datos, junto con otros predictores moderadamente fuertes. Luego, en la colección de árboles construidos, la mayoría o todos los árboles usarán este predictor fuerte en la división superior. Entonces, todos los árboles construidos a partir de *Bagging* serán similares y estarán correlacionados y la varianza difícilmente se reduzca. Los Bosques Aleatorios superan este problema obligando en cada división a considerar solo un subconjunto de los predictores. Por lo tanto, en promedio $\frac{(p-m)}{p}$ de las divisiones ni siquiera considerará el predictor fuerte, por lo que otros predictores tendrán más posibilidades de ser utilizados. Podemos pensar en este proceso como *decorrelacionar* los árboles, lo que hace que el promedio de los árboles resultantes sea menos variable y, por lo tanto, más confiable.

Por último, es importante comentar dos medidas que se derivan del método *Bagging* (incluido *Random Forest*). La primera es el error *fuera de la bolsa* (denominado en inglés *Out of bag error* u OOB) y, la segunda, es la Importancia de Variables (*Variable Importance*).

Como se mencionó anteriormente, con *Bagging* se realizan submuestras mediante *Bootstrap* y se construyen árboles con las mismas. Se puede demostrar que, en promedio, cada árbol construido utiliza alrededor de dos tercios de las observaciones (James et al., 2013). El tercio remanente que denominamos observaciones *fuera de la bolsa*, son utilizadas para predecir con los árboles entrenados. Es posible promediar (en caso de regresión) o tomar el voto mayoritario (en caso de clasificación) las predicciones realizadas en las observaciones *fuera de la bolsa*. Lo anterior lleva a una única predicción *fuera de la bolsa* y así, es posible obtener un ECM (en caso de regresión) o el error de clasificación. Estas medidas de error son válidas ya que se realizan utilizando muestras que no se utilizan para construir el árbol.

La Importancia de las Variables es un concepto útil que ayuda a la interpretación de estos métodos agregativos. También es útil para seleccionar variables cuando se cuenta con una gran cantidad y se procura eficiencia en la utilización de las mismas. Si una de las ventajas de los árboles de clasificación y regresión era su interpretabilidad y transparencia, en el caso de Bosques Aleatorios ya no es posible comunicar resultados de forma tan clara. Como se ilustró anteriormente se gana precisión predictiva/flexibilidad a costa de interpretabilidad.

Siguiendo a Genuer y Poggi (2017), la importancia de una variable se define por el aumento promedio del error (ECM) de un árbol en el bosque cuando los valores observados de esta variable se permutan aleatoriamente en las muestras OOB, respecto al mismo árbol sin perturbar. Si el error aumenta mucho ante esta permutación aleatoria *sin sentido* respecto a la que no se perturba, significa que la variable es muy importante.

Se define el índice de la importancia de X_j de la siguiente manera:

$$VI(X_j) = \frac{1}{q}\sum_{l=1}^{q}(\widetilde{errOOB}_j^{\,l} - errOOB^l) \tag{8}$$

Donde X_j es uno de los p predictores, q es la cantidad de árboles que forman el bosque (en general 500), l es la muestra seleccionada a través de *Bootstrap* (por lo que OOB^l refiere a la muestra *Out of bag* l). El término $errOOB^l$ es el error OOB (ECM en caso de regresión, proporción de mal clasificados de lo contrario) y $\widetilde{errOOB}_j^{\,l}$ es el error cuadrático medio en el caso que se permutó aleatoriamente la variable j en la muestra l. De esta forma, cuanto mayor es el error ante la permutación sin sentido, más importante es la variable.

3.3. ***Estrategia metodológica***

El objetivo inicial del trabajo es comparar tres modelos en relación a su poder predictivo. En primer lugar, se delimita el trabajo al departamento de Montevideo, por ser el área geográfica de mayor cantidad de datos y, en el caso de los datos de transacciones, mayor confiabilidad de los datos de catastro (Landaberry y Tubio, 2015). La estrategia de comparación elegida consiste en considerar pocas variables explicativas en ambas bases de datos. En el caso del modelo de ofertas las variables son: superficie construida (metros cuadrados), distancia a la playa (construida a partir de datos geográficos, en kilómetros), barrios agrupados (agrupación de barrios por cercanía), condición (usado o nuevo), dormitorios (como factor), cantidad de baños, ascensores (solamente en el caso de apartamentos) y garage. La elección de las variables se basa en la confiabilidad y en la baja presencia de datos faltantes.

En el caso de las transacciones, las variables seleccionadas son: superficie construida y del terreno (metros cuadrados), distancia a la playa (variable construida a partir de datos geográficos, en kilómetros), barrios agrupados (agrupación de barrios por cercanía), estado según catastro (variable que vale 1 si está en muy buen estado, 2 en estado normal y 3 en estado regular y malo), categoría (refiere al tipo de construcción: económica, normal, confortable, etc.), garage y patio. En este caso, la confiabilidad de los datos es alta, principalmente para propiedad horizontal (asimilable a apartamentos), aunque la disponibilidad de variables es más restringida respecto a las ofertas.

La modelización se realiza en cada base de datos por separado y por tipo de propiedad (apartamentos y casas). La separación entre casas y apartamentos se realiza por constatar que se trata de bienes muy diferentes, distribuidos geográficamente de forma muy desigual, con diferentes características (por ejemplo, la cantidad de casas nuevas es muy baja, en general las casas son mucho más grandes que los apartamentos, más antiguas, etc.). Además, en el caso de las transacciones para las que se toman datos

catastrales, la naturaleza jurídica de la propiedad horizontal hacen que la calidad de los datos sea considerablemente mayor que en el régimen de propiedad común (casas).

A los efectos expositivos, consideraremos como apartamento a la propiedad horizontal y como casa a la propiedad común. Las cuatro bases de datos a considerar son ofertas de apartamentos, ofertas de casas, transacciones de apartamentos y transacciones de casas.

En primer lugar, para cada una de las cuatro bases de datos se realiza una partición aleatoria con el paquete CARET (Kuhn y otros., 2018) para generar una muestra de entrenamiento (*training*) con el 70 % de los datos y otra de testeo (*testing*) con el restante 30 %. Los porcentajes de la partición son los más recurrentes en la literatura. A la hora de particionar las bases es necesario corroborar que ambas muestras tengan suficientes datos y se distribuyan de manera similar.

Para comparar los modelos, se entrena cada modelo realizando validación cruzada tomando cinco *folds* (*k-folds*, $k = 5$) y tres repeticiones [3]. La elección de estos parámetros es la básica sugerida en la literatura y es coherente con la cantidad de datos y el poder de procesamiento de la computadora en que se trabaja. Las particiones utilizadas en cada validación cruzada son las mismas para cada modelo (se utiliza la misma *semilla* antes de correr cada modelo). Para el caso del modelo de regresión lineal estimado por MCO, la validación cruzada solamente prueba la inclusión de la constante o no. Para este modelo se realiza un diagnóstico simple para detectar potenciales problemas (significación del modelo en su conjunto, normalidad de los residuos, presencia de datos atípicos, entre otros). También se chequea el factor de inflación de la varianza (FIV) para detectar problemas de multicolinealidad (al igual que se hace en Čeh et al. (2018)). Considerando la regla comentada en James et al. (2013), valores de FIV mayores a 10 indican una multicolinealidad problemática entre predictores (varianza excesiva de los $\hat{\beta}$). En el caso del modelo de árbol el parámetro que selecciona la validación cruzada es α y en bosques aleatorios m.

Finalmente, basado en los resultados de la validación cruzada, se toman las medidas de performance en la muestra de testeo (el 30 % reservado en la primera partición) considerando cada modelo que emerge de la validación cruzada. Una vez seleccionado el mejor de ellos, se intenta incorporar todas las variables en ese modelo y estimarlo nuevamente para lograr un mejor poder predictivo en la base de datos de testeo.

3.3.1. Medidas de performance

La mejor performance se tomará considerando medidas que serán evaluadas en la base de datos de testeo como se definió en la sección anterior. Independientemente de la validación cruzada dentro de la muestra de entrenamiento, la elección del mejor

[3] Un detalle de este procedimiento en el software R puede encontrarse en Kuhn (2018).

modelo tomará en cuenta la menor Raíz del Error Cuadrático Medio (RECM) y el menor Error Porcentual Absoluto Medio (EPAM).

Se define el RECM de la siguiente manera:

$$RECM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{9}$$

Donde N es la cantidad de observaciones en la base correspondiente (de testeo en este caso), y_i es el precio efectivo del inmueble en la base de testeo y $\hat{y}_i$ es el precio predicho en la base de testeo a partir del modelo construido con datos de entrenamiento.

Se define el EPAM de la siguiente manera:

$$EPAM = \sum_{i=1}^{N}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right| \tag{10}$$

En el caso que se proponga un problema de clasificación, la métrica del error será el porcentaje de predicciones correctas sobre el total de observaciones en una muestra de testeo.

4. Datos y análisis exploratorio

Como se mencionó anteriormente, la primera selección de datos se realiza para tomar apartamentos y casas para la ciudad de Montevideo. Se trata de una ciudad de más de 1.300.000 de habitantes (INE, Censo de 2011), que cuenta con un puerto natural, una extensa costa con playas y una gran cantidad de espacios verdes. La densidad de población es alta y alcanza los 2.500 habitantes por kilómetro cuadrado[4], la mayor densidad de población se observa en las zonas céntricas y costeras[5]. La cantidad de viviendas, según datos de catastro, supera las 500.000. Las transacciones de compra-venta de vivienda en los últimos años, según datos de la DGR, promedian las 14.000 por año (flujo de compra-ventas) mientras que las ofertas disponibles en el sitio mercadolibre.com de apartamentos y casas en venta totalizaron 80.000 publicaciones únicas para el período febrero 2018 - enero 2019.

Las bases de datos se dividen en dos tipos: una correspondiente a ofertas y la otra a transacciones. Los datos de oferta corresponden en esta oportunidad a publicaciones del sitio mercadolibre.com. En el caso de las transacciones, se utilizan registros adminis-

[4]Esto lo diferencia respecto al promedio de 19 habitantes por km^2 en todo el país.
[5]Una explicación de los datos del censo de 2011 para Montevideo puede verse en IMM (2013)

trativos recopilados por la oficina pública que tiene como principal cometido certificar el cambio de propiedad de los inmuebles.

En las secciones siguientes se presenta cada base de datos y se visualizan algunas de las principales variables. Los paquetes de R utilizados para ello son principalmente *ggplot2* (Wickham, 2016) y *leaflet* (Cheng et al., 2018).

4.1. ***Datos de ofertas***

Los datos de ofertas fueron recopilados a través de la API (interfaz para acceder a la página web) puesta a disposición por mercadolibre.com. Para ello se utilizó un programa elaborado en *python*, utilizando la biblioteca *Beautiful Soup* (Richardson, 2007). Esto sería posible lograrlo a través de R también, utilizando paquetes como *rvest* (Wickham, 2019), *httr* (Wickham, 2018), *jsonlite* (Ooms, 2014), entre otros. La muestra original incluye todas las ofertas de venta de inmuebles para la ciudad de Montevideo para el período febrero 2018 - enero de 2019 inclusive. Se realizaron bajadas sucesivas de datos alrededor del día 25 de cada mes [6].

A grandes rasgos, se eliminaron duplicados totales (publicaciones que referían al mismo inmueble), se recodificaron variables para hacerlas coherentes y se eliminaron datos erróneos y sin sentido (incluida la eliminación de valores extremos de algunas variables). También, se redujeron los niveles de las variables de tipo factor. En el caso de la variable precio y superficie construida, se optó por eliminar los extremos. Se eliminaron valores incoherentes que mostraban sucesiones de números así como números con dígitos repetidos (por ejemplo 11111, 9999, 12345, etc.).

Luego de la limpieza, la base de datos cuenta con aproximadamente 90.000 observaciones únicas (inmuebles cuyo ID no se repite, excepto que haya cambiado de precio en algún momento, en cuyo caso se opta por mantener ambos datos). 80.000 corresponden a publicaciones con un único ID, el resto refiere a las que cambiaron de precio.

La cantidad de variables son 80 aproximadamente; entre las que se encuentran el precio, título, fecha de publicación, mes de recopilación, tipo de inmueble (apartamento o casa), ubicación (coordenadas geográficas y dirección), superficie del inmueble y del terreno, cantidad de baños, de dormitorios, etc. También se cuenta con diversas variables de *amenities*: la mayoría de las mismas indican si se tiene la característica o NA. Dada la calidad optativa del dato, puede existir un sesgo en las variables NA: en algunos casos no se sabe si significa que no tiene la característica o si simplemente no se indica.

A continuación se expone una selección de visualizaciones descriptivas de los datos para explorar la relación entre la variable de interés (el precio) con variables explicativas.

En primera instancia, observamos la cantidad de publicaciones de apartamentos y casas recopiladas mes a mes a través de la Api del sitio en la Figura 2.

[6] Los datos de oferta están a disposición aquí.

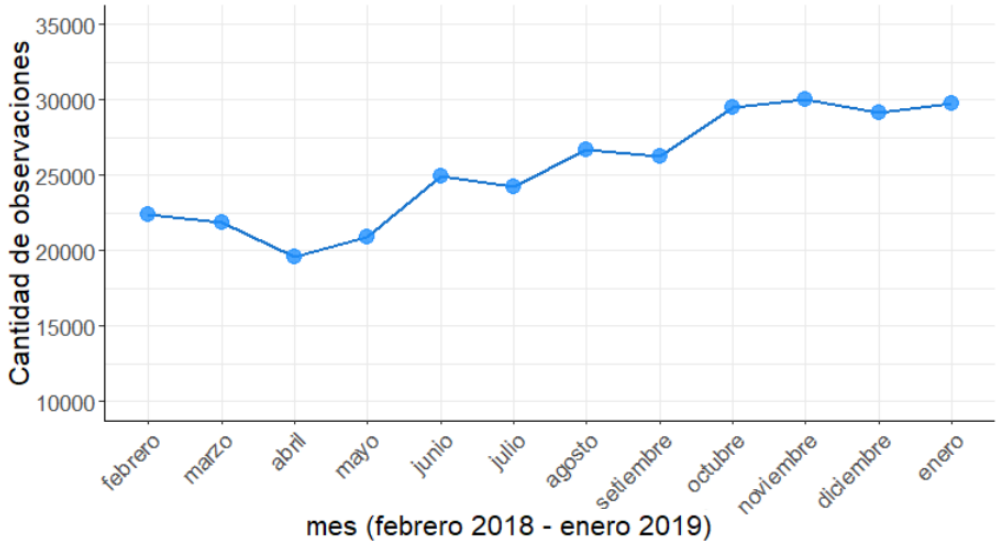


Figura 2: Cantidad de publicaciones recopiladas de apartamentos/casas por mes. Refiere a la cantidad que existe en el momento de la bajada de datos (stock). Fuente: Elaboración propia, en base a mercadolibre.com

Los datos parecerían indicar un crecimiento de las ofertas aunque el comportamiento podría asociarse a estacionalidad. Si bien no se exponen en este trabajo, las publicaciones de febrero a agosto de 2019 (mes a mes) fueron recopilados y superan las 30.000.

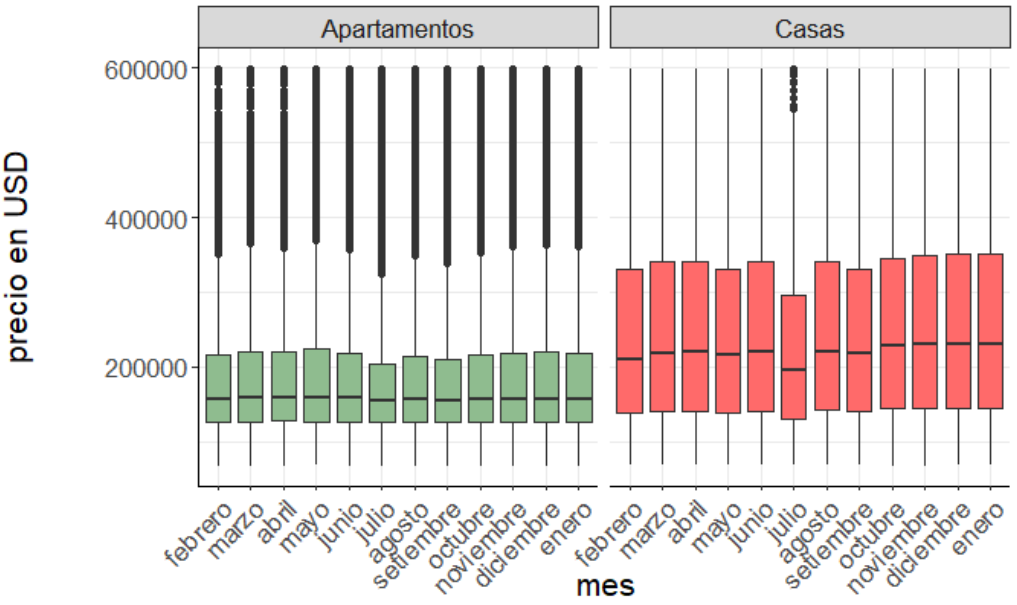


Figura 3: Precio durante 2018 para apartamentos/casas por mes (incluye ene-19) Fuente: Elaboración propia, en base a mercadolibre.com

Respecto a los precios en dólares estadounidenses (USD corrientes) durante el período en que se recopilaron los datos, no hay cambios abruptos en la distribución, salvo en el

mes de julio de 2018 que se observa una caída circunstancial (Figura 3).

Como se mencionó en la introducción, la dolarización es total en el mercado inmobiliario uruguayo y la apreciación repentina que experimentó la moneda estadounidense en mayo y setiembre de 2018 no parece haber afectado de forma notoria a los precios de las ofertas.

En relación al tipo de inmueble, los apartamentos dominan las ofertas (las transacciones también). Si tomamos la distribución por barrios, se destacan Pocitos, Cordón, Centro, Malvín y Punta Carretas que concentran más del 50 % de las ofertas recopiladas y la gran mayoría refieren a apartamentos (Figura 4). El caso del barrio Carrasco es atípico pues, siendo un barrio con gran cantidad de ofertas, las de casas superan a las de apartamentos.

Respecto a la distribución según la condición de los inmuebles predominan los usados. Se destaca que esta condición abarca casi la totalidad de las casas. La proporción de apartamentos nuevos asciende a 1/4 de ese tipo de propiedad.

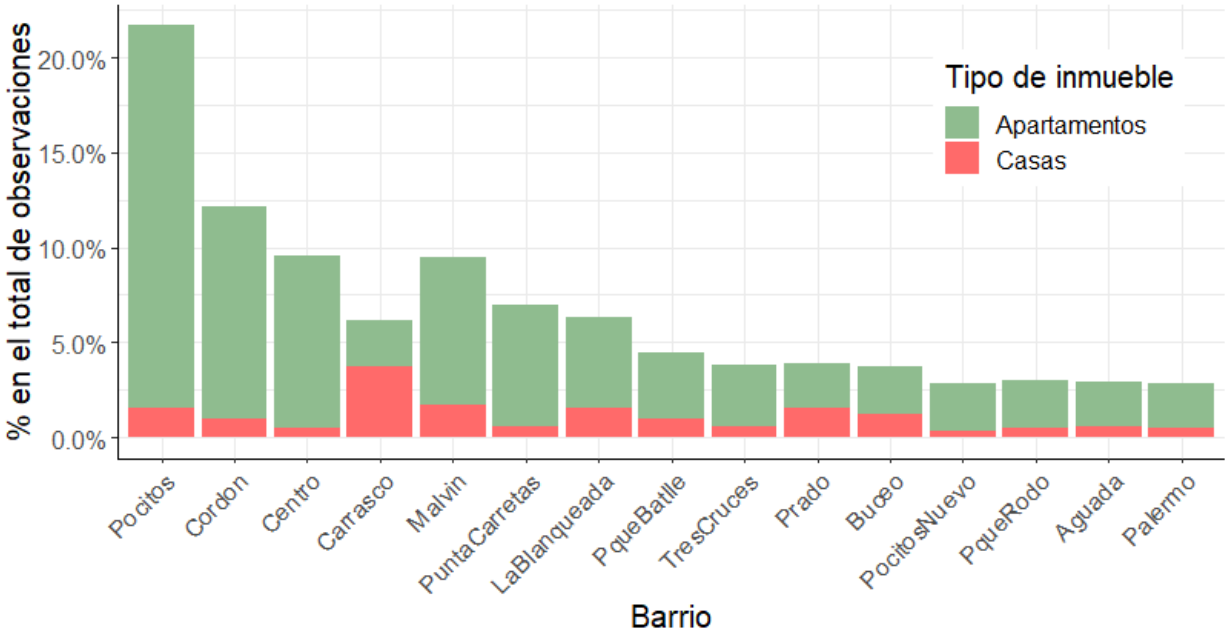


Figura 4: Observaciones de apartamentos y casas por barrios
Barrios con más de 2000 observaciones
Fuente: Elaboración propia, en base a mercadolibre.com

La distribución de la variable precio, tiene una cola larga derecha incluso luego de filtrar datos extremos como puede observarse en la Figura 5. Este comportamiento es más marcado en el caso de los apartamentos. No obstante, en las casas es donde se presenta mayor variabilidad de los precios (considerando el rango intercuartílico).

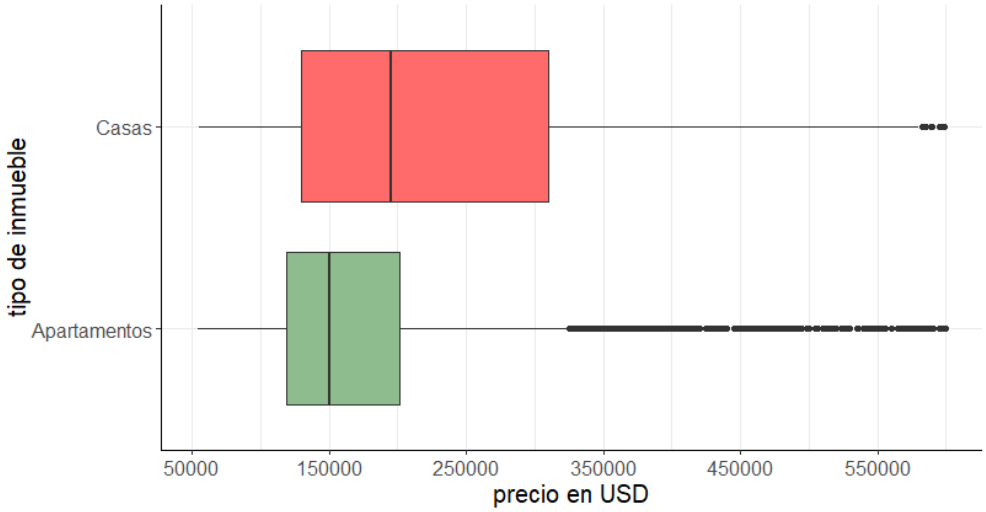


Figura 5: Distribución de la variable precio por tipo de inmueble
Fuente: Elaboración propia, en base a mercadolibre.com

Si en vez de considerar el precio total, tomamos en cuenta el precio por m^2 se observa una *normalización* de la distribución de los precios, en el caso de los apartamentos con una media en torno a USD 2500. No sucede lo mismo con las casas, donde persisten precios a la derecha de la distribución y una media de 1700 USD (Ver Figura B.25 en el Apéndice B).

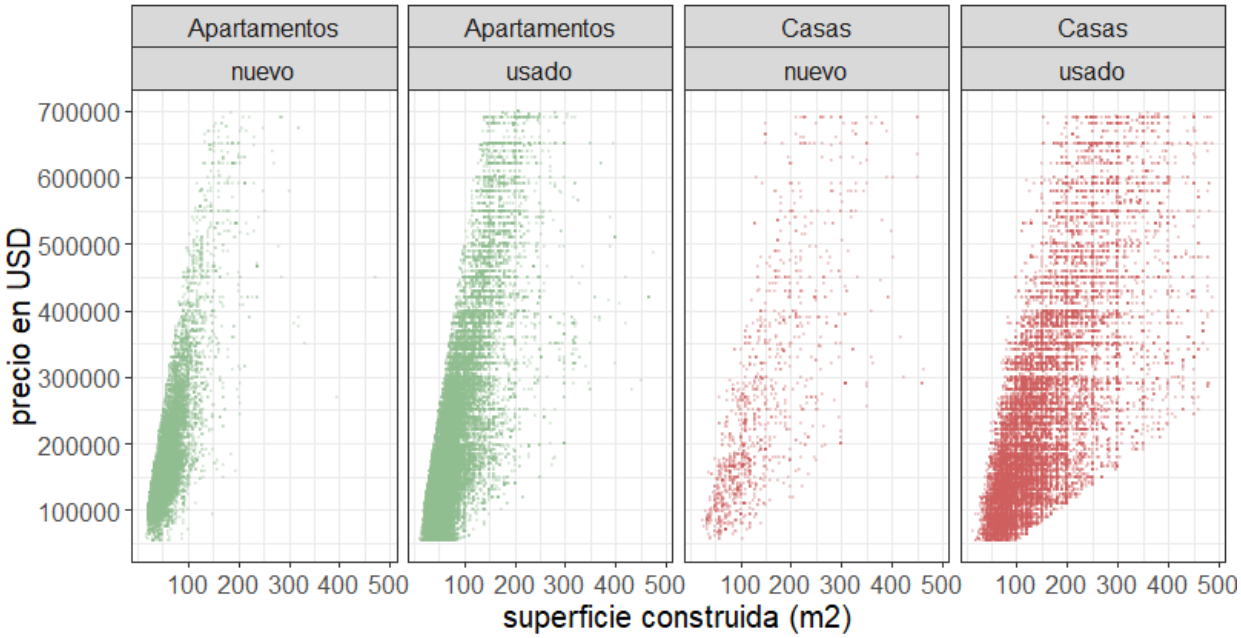


Figura 6: Dispersión del precio y superficie construida (m^2) por tipo y condición
Fuente: Elaboración propia, en base a mercadolibre.com.

La relación entre el precio y la superficie construida es claramente creciente, principalmente para los apartamentos que en su mayoría se concentran en los que tienen una superficie menor a 100 m^2. En las casas, la dispersión de precios y superficie es mayor, lo que refuerza la idea de que se trata de bienes más heterogéneos que los apartamentos (Figura 6).

Esta relación se desdibuja al considerar el precio por m^2 (Figura B.24 en Apéndice B). En referencia al precio y al precio por m^2 según tipo de propiedad y barrio, existen diferencias por barrio. En Carrasco, por ejemplo, se encuentran las propiedades más caras considerando la mediana del precio total. Sin embargo, si consideramos el precio por m^2, el barrio más cotizado es Punta Carretas. Lo anterior se ve influido por la mayor presencia de propiedades grandes en Carrasco y el efecto del tamaño en el precio del m^2 (negativo).

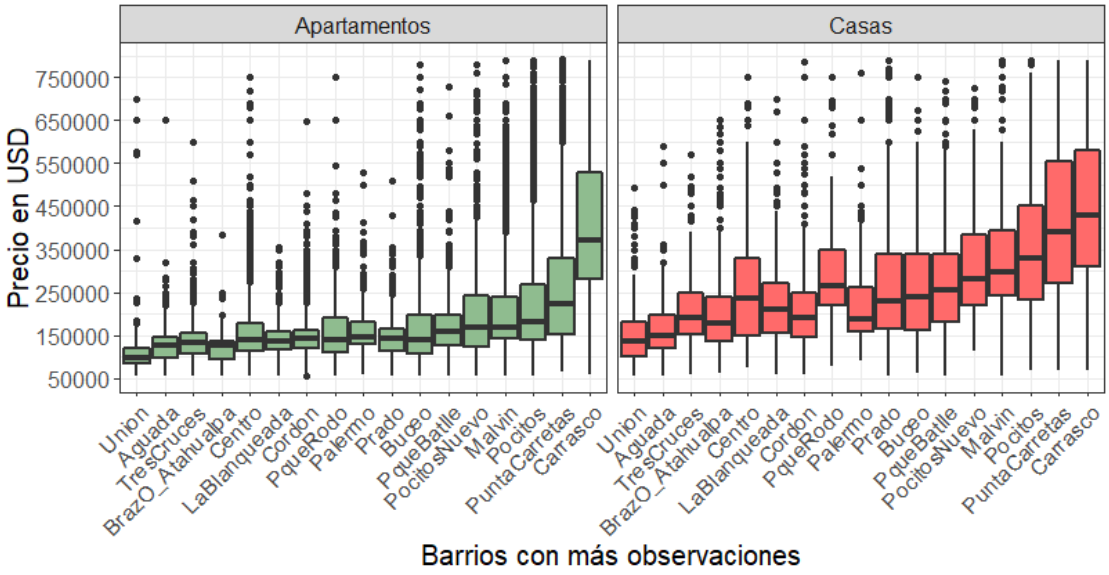


Figura 7: Distribución del precio por barrios y tipo de propiedad
Fuente: Elaboración propia, en base a mercadolibre.com.

4.1.1. Ubicación de las ofertas

Dado que uno de los datos indicados en la publicación del inmueble en mercadolibre.com es la dirección (y se trata de un atributo que es *obligatorio* indicar), es posible deducir las coordenadas geográficas (latitud y longitud). Se procuró chequear ubicaciones repetidas muchas veces pues podrían corresponder a la dirección de una inmobiliaria u otro. A continuación se presentan dos visualizaciones geográficas del precio de apartamentos según rangos de precios (delimitados por los quintiles y tipo de inmueble).

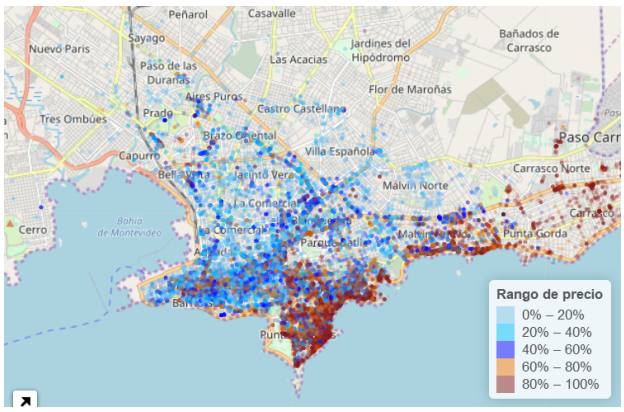


Figura 8: Ubicación de ofertas de aptos

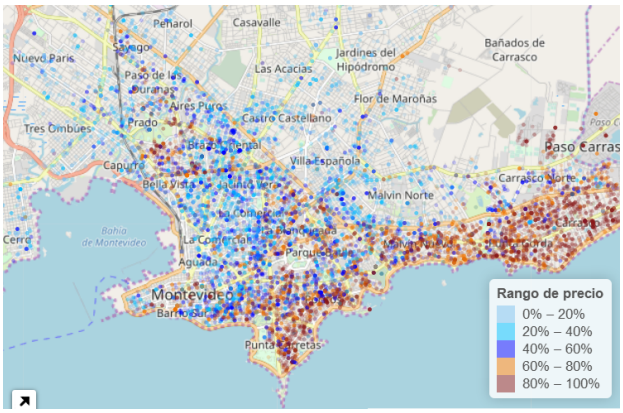


Figura 9: Ubicación de ofertas de casas

Rango de precios en base a los quintiles por tipo de propiedad.
Fuente: Elaboración propia, en base a DGR, DNC, SIG-IM

Se observa un hecho estilizado claro: propiedades más caras se ubican sobre la costa este de la ciudad. Esto se refuerza cuando se observa el precio del m^2. Hay algunas excepciones a este hecho: el barrio residencial del Prado y alrededores (Figuras 8 y 9).

Se puede apreciar que la distribución de las casas se distribuye de forma más extensiva en la ciudad respecto a los apartamentos. Por ejemplo, en la zona oeste, precisamente en los barrios Cerro, Tres Ombúes, Nuevo París, Sayago, etc., figuran muy pocos apartamentos, no sucede así con las casas. Si se considera el precio por m^2 de los inmuebles, se confirma aún más la idea del alto precio en la costa este de la ciudad (Figura C.27 en Apéndice C).

4.2. ***Datos de transacciones***

La base de datos de transacciones abarca el período enero 2017 - junio de 2018 inclusive e identifica transacciones de compraventa de padrones en Montevideo con destino de vivienda. Los datos fueron otorgados por la Dirección General de Registros (DGR). Lo más novedoso de la base de datos, además de corresponder a compra-ventas efectivamente realizadas, es la posibilidad de identificar la ubicación exacta del inmueble. Esto se realiza a través del código de padrón y la unión con bases de datos geográficos provistas por la Intendencia de Montevideo (IM), así como la unión con características disponibles a partir de la Dirección Nacional de Catastro (DNC). Estas dos últimas bases de datos son de libre disposición.

Para conformar la base final de transacciones fue necesaria la unión entre la base de transacciones (de DGR) con la de información geográfica (del Sistema de Información Geográfica de la Intendencia de Montevideo) y con la información catastral (DNC).

Primero se cargaron los datos del parcelario de Montevideo y se proyectó en el sistema de coordenadas estándar que es compatible con los datos que se toman de ofertas

y con los mapas que se utilizan (Leaflet, Google maps, etc.). La base de información geográfica es un archivo *shape* (shp) con los padrones delimitados en forma de polígonos. Para obtener una sola coordenada (latitud y longitud) por padrón, fue necesario transformar las coordenadas de un polígono en las de un punto. Para lograr lo anterior se calcularon *centroides* para cada padrón: posición media aritmética de todos los puntos del polígono.

En una segunda instancia, se unieron de los datos catastrales, con los *centroides* calculados, con la información geográfica y los datos de transacciones.

Una vez generada la base de datos a partir de las tres fuentes, se realiza el procesamiento de limpieza. La identificación de cada transacción consta del número de padrón, unidad/block (si corresponde) y fecha de transacción. Notar que una misma propiedad puede presentarse varias veces (transacciones sobre el mismo inmueble en diferentes momentos).

Un vez conformada la base de datos, se tomaron las transacciones en USD (más del 95 % de la muestra total), pues las denominadas en otras monedas podrían no representar operaciones *de mercado*. Se eliminaron observaciones con valores extremos en las variables más relevantes (precio, precio del m^2, superficie construida y del terreno) y se agruparon observaciones en barrios por cercanía.

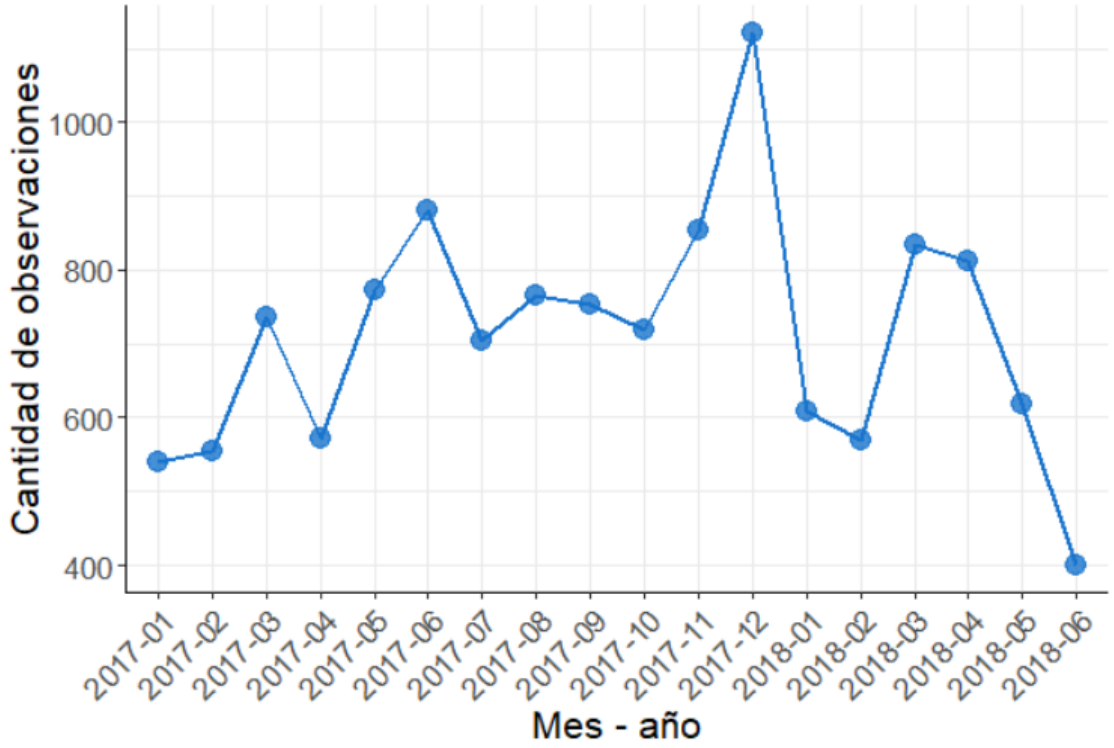


Figura 10: Cantidad de transacciones por mes (en base de datos procesada)
Fuente: Elaboración propia, en base a DGR y DNC

La muestra seleccionada y procesada cuenta con 12.815 observaciones que presentan cierta estabilidad en la cantidad por mes, como se puede observar en la Figura 10 (excepto en diciembre de 2017 y junio de 2018, que hay un máximo y mínimo respectivamente).

Luego de remover extremos y procesar los datos, la distribución de los precios por mes muestra estabilidad, como se observa en la Figura 11. La distribución de las transacciones según el tipo de propiedad es similar a las ofertas (80 % de apartamentos y 20 % de casas). La antigüedad de los inmuebles se presenta en forma de factor, identificando cinco grandes grupos con cantidades de transacciones similares.

En otro orden, la información de la base de datos geográfica incluye una variable denominada *área diferencial* (asociado al barrio, en algunos casos, de manera agrupada).

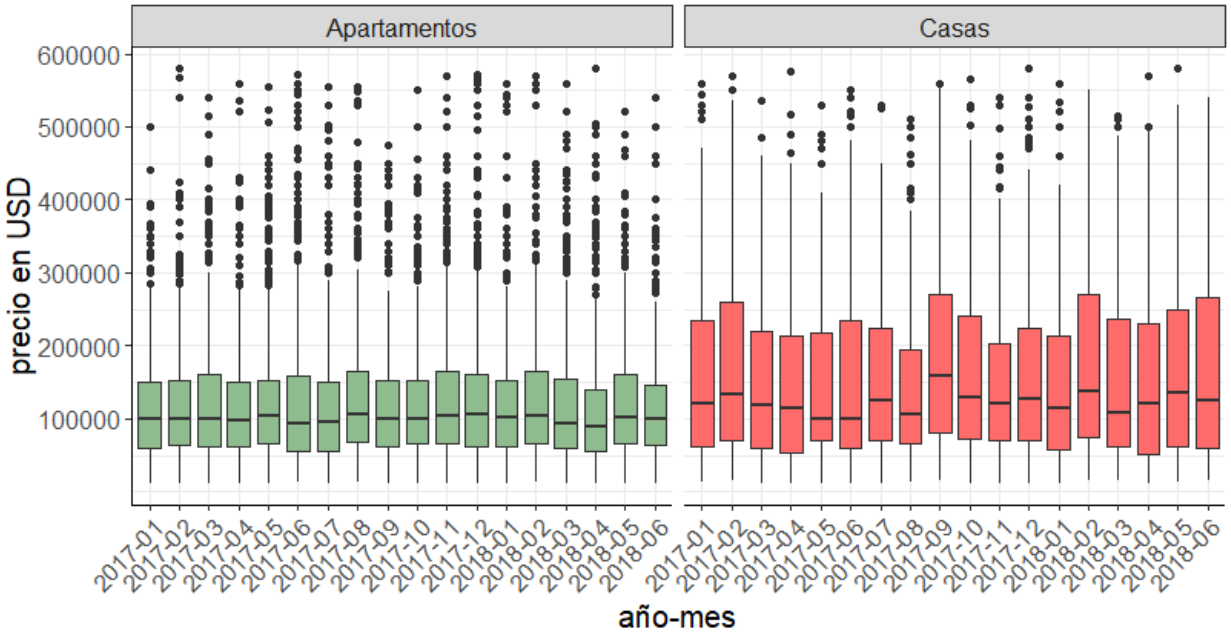


Figura 11: Precio en USD de las transacciones durante enero 2017 - julio 2018 por tipo de inmueble (apartamentos y casas)
Fuente: Elaboración propia, en base a DGR y DNC

Como se desprende de la Figura 12 y a diferencia de las ofertas, hay una distribución más equitativa de los barrios en la muestra. Por ejemplo, las áreas Prado-Capurro, Jacinto Vera-La Blanqueada y La Comercial-Villa Muñoz aumentan su peso relativo.

Se destaca el Centro, que al incluir a Cordón, es el barrio con mayor cantidad de transacciones, superando a Pocitos; el barrio más densamente poblado y con mayor cantidad de ofertas (incluso comparando con las ofertas de Centro-Cordón).

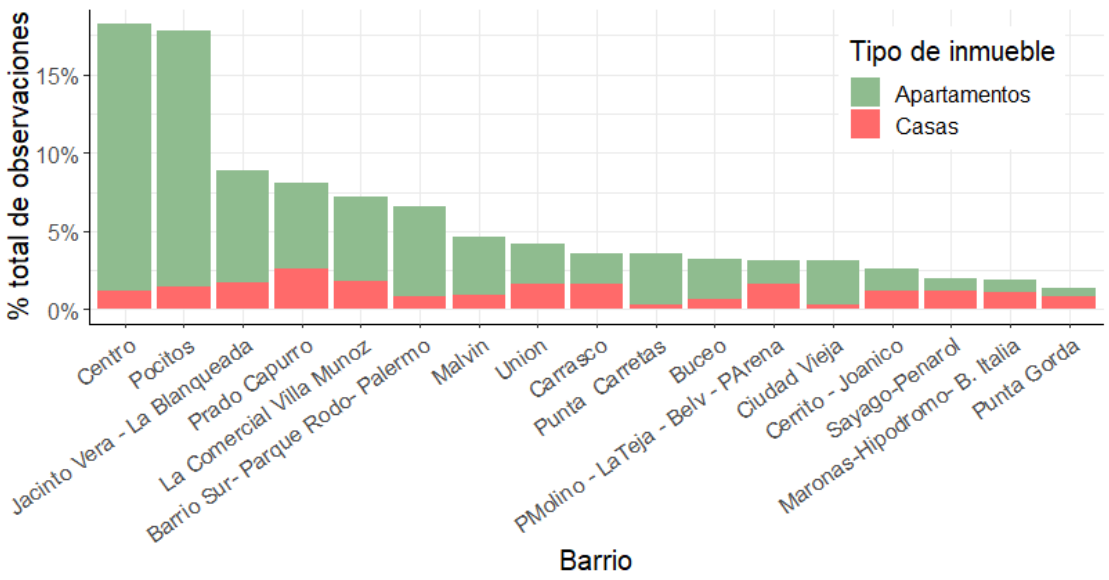


Figura 12: Cantidad de transacciones por tipo de propiedad (apartamentos o casas) y por barrio (área diferencial)
Fuente: Elaboración propia, en base a DGR y DNC

La distribución de los precios según tipo de inmueble es similar a las ofertas; mayor variabilidad en el precio de las casas (se invierte al considerar m^2, ver Apéndice B).

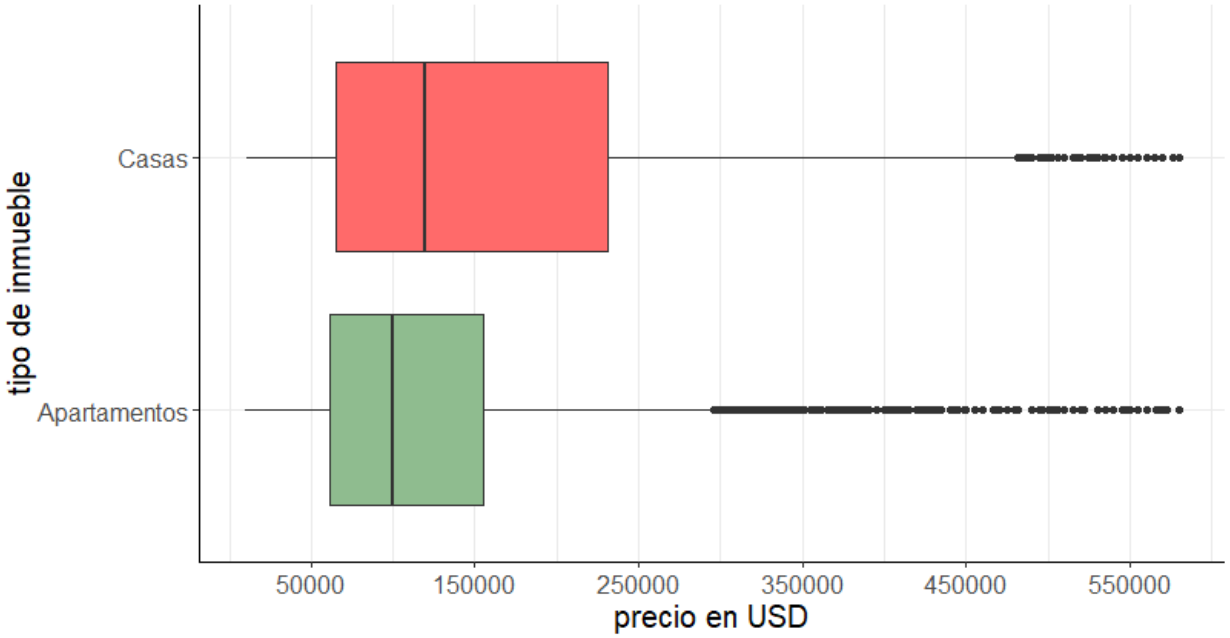


Figura 13: Distribución del precio para apartamentos y casas
Fuente: Elaboración propia, en base a DGR y DNC

La distribución del precio por barrio (Figura 14), presenta un patrón similar a las ofertas: los precios son mayores sobre la costa este, más aún tomando el precio por m^2.

Por último, existe una variable tomada de la DNC que refiere al valor catastral del inmueble. La misma está en pesos uruguayos y su objetivo es establecer una referencia para ciertos impuestos (como porcentaje del valor catastral). Si bien este valor no corresponde al valor de mercado, sí existe una relación entre ambos. Tomando el promedio anual del tipo de cambio, la relación para el 50 % central de los datos entre el valor catastral y el precio de mercado es de entre 2 y 3,8 veces. Si bien se observa una correlación lineal positiva, se destacan desvíos que no explicados totalmente por las variables con las que se cuenta.

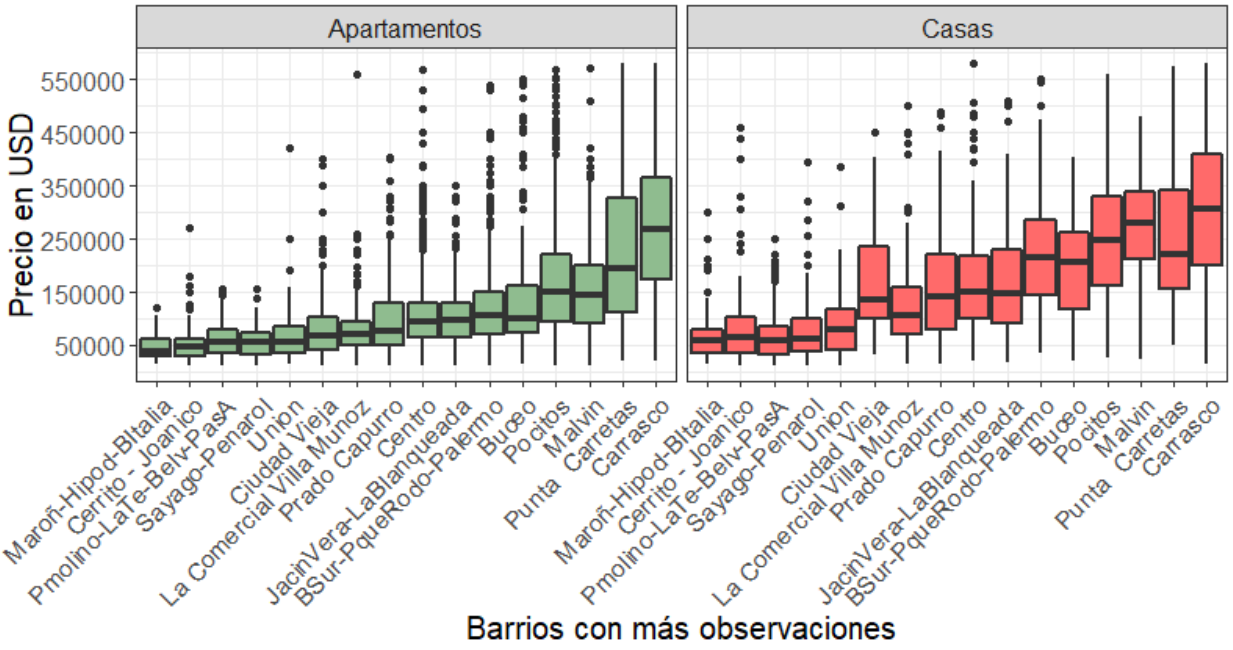


Figura 14: Precio por barrios y tipo de propiedad
Fuente: Elaboración propia, en base a DGR y DNC

En la Figura 15, se visualiza la relación entre el valor catastral y de mercado, conjuntamente con el estado de conservación del inmueble (a partir del catastro). Se puede apreciar que propiedades en excelente estado tienen menor dispersión entre el valor catastral y el de mercado. Para el resto de los inmuebles existe una relación menos clara.

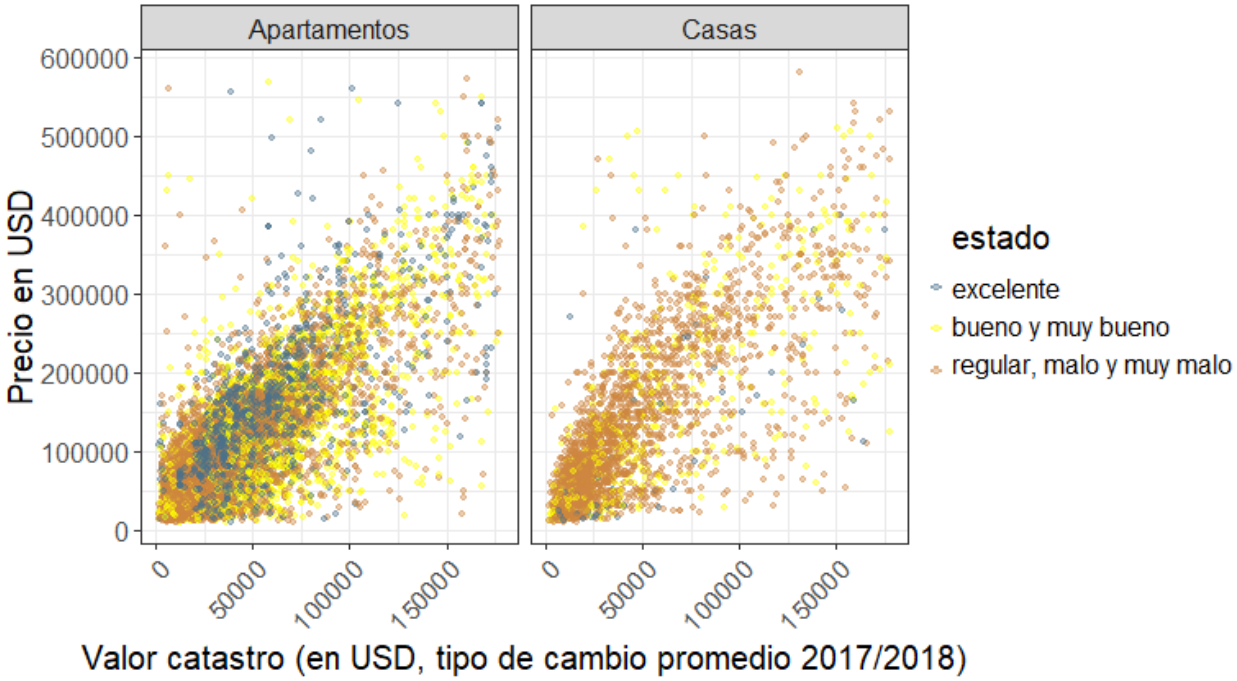


Figura 15: Precio transado (en USD), valor catastral (en USD) y estado de conservación
Fuente: Elaboración propia, en base a DGR y DNC

4.2.1. Ubicación de las transacciones

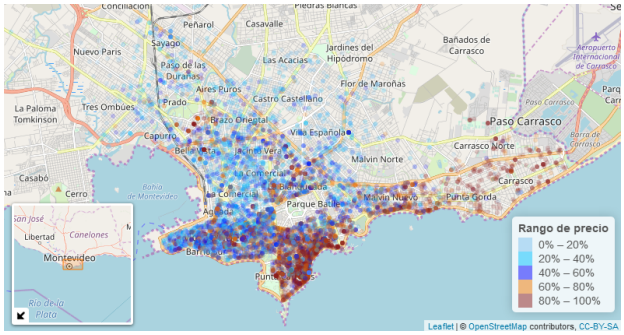


Figura 16: Ubicación de Apartamentos transados y rangos de precio

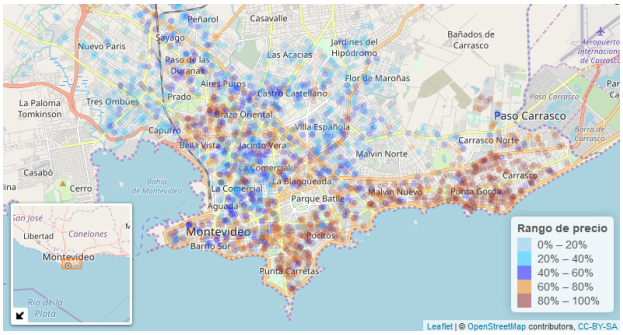


Figura 17: Ubicación de casas transadas y rango de precios
Rango de precios en base a los quintiles por tipo de propiedad.
Fuente: Elaboración propia, en base a DGR, DNC, SIG-IM

El patrón de precios según la ubicación se confirma también para las transacciones. Se destaca que las propiedades transadas se encuentran más dispersas geográficamente, principalmente las correspondientes a casas.

5. Precios de oferta vs transacción

El contraste entre las variables de precio en ambas bases puede dar lugar a diferencias relevantes. En la literatura se aborda el tema como brecha entre el precio pedido y el precio efectivo. Esta brecha puede tener comportamientos diferentes según la coyuntura, el ciclo económico y/o la tensión del mercado inmobiliario (Haurin et al., 2013, Horowitz, 1992, Carrillo, 2013). A su vez, el precio pedido puede tener un rol como predictor, guía o cota del precio final (Benítez-Silva et al., 2015, Han y Strange, 2016, Dubin, 1998). En la mayoría de los trabajos se enfatiza el análisis conjunto de ambos precios, la magnitud de la brecha y su signo.

Para realizar la comparación es importante controlar por características de los inmuebles en las dos bases de datos. Los datos de ofertas tienen sesgos de diversa índole (por tipo de propiedad, por barrio, entre otros) que hacen que no sean representativos de todos los inmuebles de la ciudad ni de los que se compran y venden. A su vez, en ambas bases de datos se tiene una muestra acotada en el tiempo, lo que afecta más a las transacciones, que son relativamente menos numerosas.

Se presenta una comparación para el barrio Pocitos (15 % de las observaciones de transacciones y 20 % de ofertas). Se toma el precio total de los apartamentos para el barrio indicado y se reporta la diferencia de precios entre las ofertas y transacciones. La ubicación de estos inmuebles se observa en las Figuras 18 y 19.

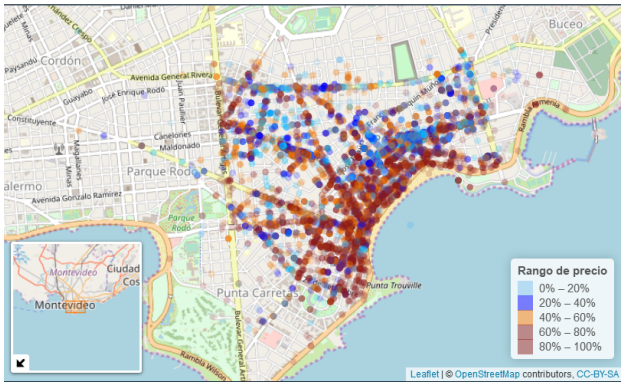


Figura 18: Ofertas de apartamentos en el barrio Pocitos

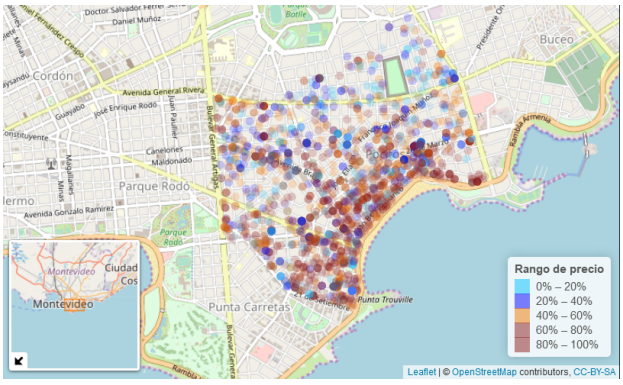


Figura 19: Transacciones de apartamentos en el barrio Pocitos

Fuente: Elaboración propia, en base a mercadolibre.com y DGR, DNC, IM.

Los precios totales ofertados se encuentran, en promedio y sin controlar por demás características un 32 % por encima de los efectivamente transados (en mediana 26 %). En el caso del precio del m^2, esta diferencia es similar (30 % en promedio y 21 % en mediana). La brecha varía según el barrio y si se controla por características, lo que será estudiado en una versión posterior de este documento de trabajo.

6. Resultados

En esta sección se presentan los principales resultados de este trabajo. En cada uno de los modelos utilizados la variable de respuesta es el logaritmo del precio del inmueble expresado en dólares estadounidenses. Tres modelos alternativos son comparados utilizando los mismos predictores: modelo lineal hedónico, árbol de regresión y bosques aleatorios. Para comparar los resultados se utiliza la validación cruzada y se evalúa la performance predictiva con base en el RECM y el EPAM. Se selecciona el mejor de los tres y se desarrolla el mismo. Este procedimiento, detallado en la Sección 3.3, se realiza con los datos de oferta y transacción por separado así como por tipo de inmueble. Se opta por dedicar más espacio a los modelos con apartamentos (de oferta y de transacción) debido a la mayor disponibilidad de datos.

6.1. ***Modelos con datos de oferta***

6.1.1. Apartamentos - ofertas

En una primera instancia se realiza la partición para obtener las muestras de entrenamiento y testeo. El 70 % de la muestra se incluye en la primera partición con 49.574 observaciones de un total de 70.817. La selección de predictores para esta instancia, consta de las siguientes ocho variables:

- *distancia a la playa*
- *barrios agrupados*
- *condición*
- *dormitorios*
- *baños*
- *log(sup. construida)*
- *garage*
- *ascensores*

6.1.2. A. Modelo lineal hedónico - ofertas de apartamentos

La estimación por MCO del modelo de regresión lineal arroja el siguiente resultado:

Cuadro 1: Modelo lineal hedónico con ofertas de apartamentos

Ver en Cuadro E.11 del Apéndice E la salida completa (incluye efectos de barrio).

	Variable dependiente: ***log(precio)***	
	coeficientes	(error std.)
log(distancia a la playa)	-0.060^{***}	(0.002)
(...)	**(...)**	**(...)**
condición usado (base nuevo)	-0.079^{***}	(0.002)
1 dormitorio (base 0)	0.060^{***}	(0.004)
2 dormitorios	0.079^{***}	(0.004)
3 o más dormitorios	0.021^{***}	(0.005)
baños (cantidad)	0.151^{***}	(0.003)
log(superficie construida)	0.615^{***}	(0.004)
garage Sí (base No)	0.110^{***}	(0.002)
constante	9.344^{***}	(0.012)
Observaciones	49,574	
R^2	0.800	
Adjusted R^2	0.800	
Residual Std. Error	0.216 (df = 49553)	
F Statistic	$9{,}907.693^{***}$ (df = 20; 49553)	
Nota:	*p<0.1; **p<0.05; ***p<0.01	

Se desprende de el Cuadro 1 que todos los coeficientes estimados son significativamente diferentes de cero (con 5 % de significación) así como el modelo en su conjunto y los signos son coherentes con lo esperado y los Antecedentes. Se destaca que un aumento de 10 % de la superficie del apartamento implica en promedio un precio de oferta 6 % mayor, dado todo lo demás constante. A su vez, alejarse de la playa disminuye el precio en promedio (0,6 % ante un 10 % más lejos). Las ofertas de apartamentos usados son en promedio un 8 % más baratas que los nuevos. A su vez, disponer de garage repercute en una oferta 11 % más cara. Los signos asociados a los barrios son coherentes, considerando que el barrio base es Pocitos.

Se calcula el FIV y no arroja indicios de multicolinealidad. Los supuestos del modelo fueron chequeados. En el caso de los residuos del modelo lineal final (Figura E.32 en Apéndice E), se constata la presencia de valores atípicos que pueden afectar al modelo aunque no son influyentes.

6.1.3. B. Árbol de regresión - ofertas de apartamentos

Con las mismas variables consideradas en el modelo anterior, se entrena el modelo de árbol de regresión. El mejor árbol seleccionado por la validación cruzada, en base al menor RECM, considera el parámetro costo-complejidad $\alpha = 0,047$.

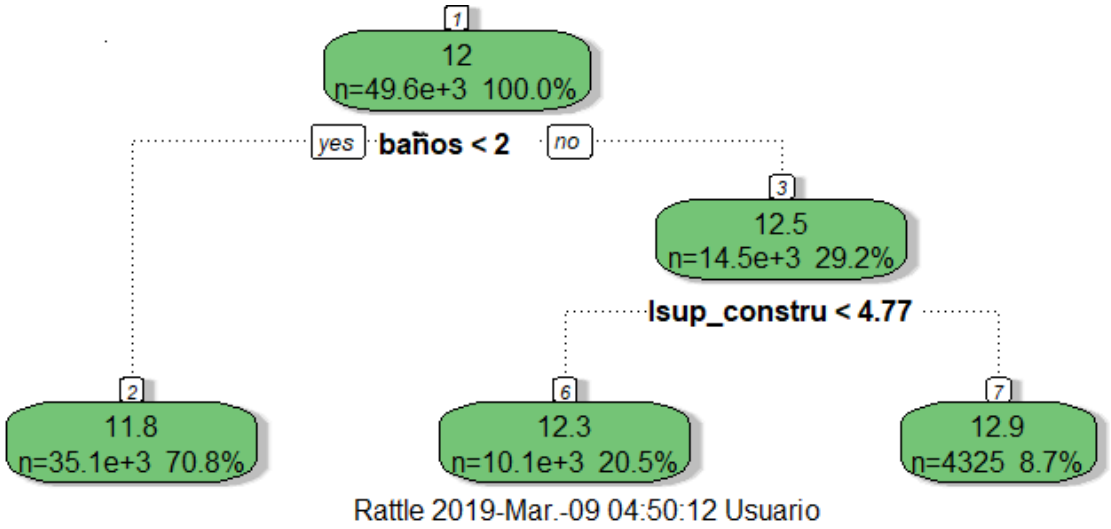


Figura 20: Árbol de regresión con ofertas de apartamentos

Se destaca la importancia de la variable baños que es la que mejor particiona la muestra. Esta variable toma el punto de corte en 2, es decir, si el apartamento tiene menos de dos baños (nodo de la izquierda), la predicción del precio del inmueble dentro de la muestra es de USD 133.200 ($e^{11,8} \approx 133{,}200$). En esta partición cae el 70,8 % de la muestra (35.100 observaciones). Por el contrario, si el apartamento tiene dos o más baños (nodo de la derecha), la predicción es USD 268.000 ($e^{12,5} \approx 268{,}000$) y contiene 14.500 observaciones (29,2 % de la muestra). El logaritmo de la superficie construida es la segunda variable utilizada para la partición y el punto de corte es aproximadamente 118 m^2 ($e^{4,77} \approx 118$). En la hoja que separa a los apartamentos con menos de 118 m^2 (y previamente con más de dos baños) se encuentran 10.100 observaciones mientras que en las que tienen más de 118 m^2, 4.325.

6.1.4. C. Bosques aleatorios - ofertas de apartamentos

Se realiza el mismo entrenamiento con bosques aleatorios, el parámetro seleccionado es $p = 10$. Este número considera cada categoría de las variables factor como una variable *dummy* (Kuhn y otros., 2018). Por ejemplo, en el caso de la variable *barrios agrupados* hay 12 categorías, por lo tanto el algoritmo de solución considera 12 variables dummy que se activan (valen 1) cuando corresponde al barrio agrupado, de lo contrario valen 0. El parámetro p es seleccionado por la validación cruzada por ser el que obtiene menor RECM de las tres repeticiones hechas.

Cuadro 2: Validación cruzada. Bosques aleatorios - ofertas de apartamentos

Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 39671, 39672, 39670, 39669, ...
Resampling results across tuning parameters:

mtry	RMSE	Rsquared	MAE
2	0.2334330	0.8012384	0.1735597
10	0.1539799	0.8989698	0.1048605
19	0.1552244	0.8971778	0.1026534

6.1.5. D. Desarrollo del mejor modelo con ofertas de apartamentos

La performance predictiva de los tres modelos anteriores indica que el modelo de bosques aleatorios logra reducir el error del modelo lineal un 40 % (de un EPAM de 17 % a 10 %) en la predicción del precio de oferta de apartamentos. A continuación se presentan los resultados.

Cuadro 3: Performance de los modelos con ofertas de apartamentos

Modelo	RECM	EPAM
Lineal	47.760	17 %
Árbol	71.420	26 %
Bosque	**30.960**	**10 %**

Los errores reportados indican que en promedio, el modelo lineal hedónico falla por 47.760 USD, o por 17 %. Mientras, el bosque aleatorio lo hace por 30.960 USD o 10 %.

Una vez demostrada la mejor performance de los bosques aleatorios en la *carrera de caballos* (*model horse-racing*), se realiza un desarrollo sin validación cruzada agregando más predictores. Se toma el parámetro $m = \sqrt{p}$, que utiliza por defecto el paquete *randomForest* (Liaw y Wiener, 2002). En este caso, se adicionan las siguientes variables (para incluir una variable de tipo factor o *dummy* al menos debe tener 2 % con valores en cada nivel):

- *barrio*
- *terraza*
- *dormitorio en suite*
- *lavadero*
- *living*
- *aire acondicionado*
- *calefacción*
- *parrillero*
- *salón comunal*
- *seguridad*
- *balcón*
- *mes*
- *expensas (gastos comunes)*
- *placard*
- *playa*
- *playa este*
- *log(distancia a la playa este)*

La incorporación de estas variables logra reducir el error en un punto porcentual. La magnitud del error debe tener en cuenta que las variables incorporadas tienen falencias. Por ejemplo, algunas de ellas son opcionales (el usuario puede indicarlas o no) y en algunos casos, fue necesario realizar imputaciones a valores NA.

Cuadro 4: Performance del bosque desarrollado. Ofertas de apartamentos

RECM	EPAM
29.530	9 %

Es posible calcular la medida de importancia de las variables, como se explicó en el Marco Teórico. La superficie construida, la condición del apartamento (nuevo/usado), garage, gastos comunes (expensas) y las asociadas a ubicación son las más relevantes.

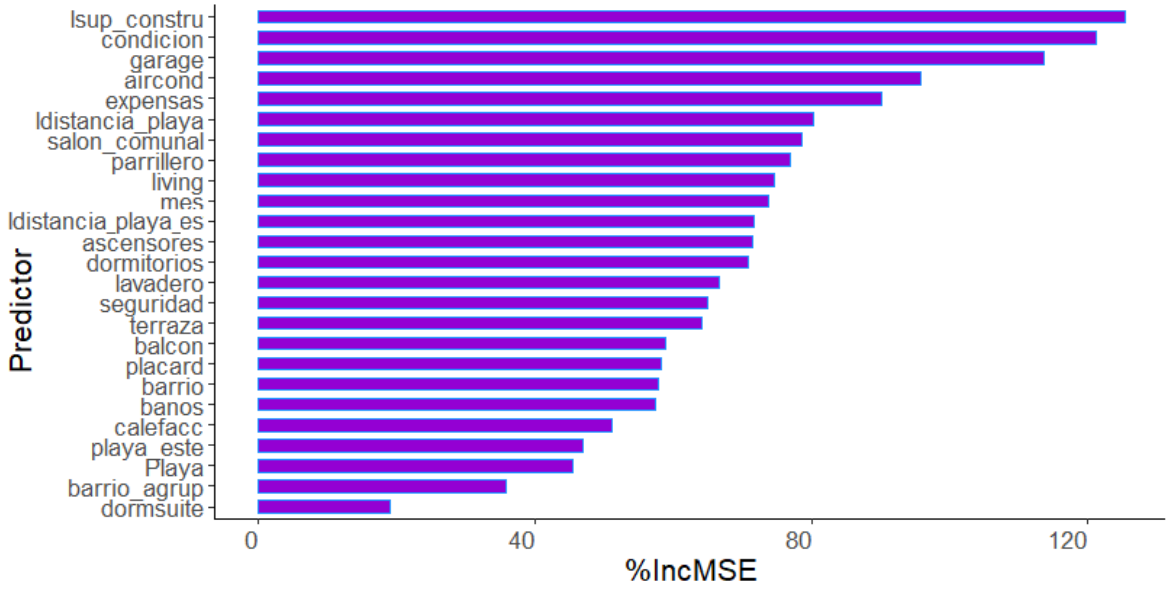


Figura 21: Medida de la importancia de las variables para predecir el precio de oferta de apartamentos - Modelo bosques aleatorios desarrollado

6.1.6. Casas - ofertas

A continuación, se realizó la misma comparación entre modelos con los datos de ofertas de casas. La partición de entrenamiento representa el 70 % de la muestra con 15.412 observaciones de un total de 22.015. La selección de predictores para esta instancia es

la misma que originalmente se realizó para apartamentos, excepto por la variable ascensores. A su vez, se modifica la variable dormitorios para agruparla en dos categorías. Por un lado, las casas con dos dormitorios o menos y, por el otro, con tres dormitorios o más. Esto se realiza porque hay pocas observaciones de casas con 0 y 1 dormitorio.

A modo de resumen, el modelo de bosques aleatorios presenta mejor performance predictiva en la comparación entre los tres modelos (Cuadro 5). Este modelo supera considerablemente al modelo lineal, con una reducción de 36 % del error porcentual absoluto medio (22 % a 14 %). Cuando se desarrolla, este error disminuye en un punto porcentual.

Cuadro 5: Performance de los modelos con ofertas de casas

Modelo	RECM	EPAM
Lineal	75.240	22 %
Árbol	126.740	29 %
Bosque	56.800	14 %
Bosque desarrollado	53.200	13 %

La predicción empeora respecto a los apartamentos para los tres casos. Esto se da principalmente porque se trata de bienes más heterogéneos y se cuenta con menos observaciones.

En relación al modelo lineal hedónico, los coeficientes estimados son significativamente diferentes de cero (al 5 %) y sus signos son coherentes, algunos con diferencias respecto a los estimados para apartamentos. Se destacan los menores efectos de *distancia a la playa*, *superficie construida*, *baños* y *garage*. Respecto al FIV, se descarta multicolinealidad.

En el árbol de regresión seleccionado con validación cruzada (ver Figura E.34 en Apéndice E) la primera partición se realiza con la superficie construida, a partir de 150 m^2 ($e^{5,02} \approx 150$). En segundo lugar, se destaca la distancia a la playa como criterio de partición, a partir de una distancia de 2,36 kms ($e^{0,86} \approx 2,36$).

El modelo de bosques aleatorios seleccionado por la validación cruzada toma un valor de $p = 17$. Con el fin de desarrollar este modelo, se agregan las mismas variables que para los apartamentos, excepto por *lavadero* (pocas observaciones que indican *Sí*) y *expensas* (no aplica). El EPAM disminuye en un punto porcentual, de forma similar a los apartamentos.

Respecto a la medida de importancia de las variables, se destaca la variable *garage* y *parrillero*, luego de *superficie construida* que resalta por el alto valor de la medida.

6.2. **Modelos con datos de transacciones**

6.2.1. *Apartamentos - transacciones*

El 70 % de la muestra de entrenamiento incluye 7.018 observaciones de un total de 10.023 apartamentos. La selección de predictores es la siguiente:

- *distancia a la playa*
- *barrios agrupados*
- *estado*
- *categoría*
- *superficie construida*
- *superficie del terreno*
- *antigüedad*
- *garage*
- *patio*

6.2.2. *A. Modelo lineal hedónico - transacciones de apartamentos*

El resultado del modelo lineal arroja que los coeficientes estimados son significativamente diferentes de cero con 95 % de confianza, excepto por el coeficiente asociado a la categoría que indica los barrios Carrasco, Punta Gorda y Malvín. Otra excepción es el coeficiente del logaritmo de la superficie del terreno, que es significativo al 90 %. Los signos de los coeficientes estimados son los esperados (recordar que la variable antigüedad tiene un rango de 1 a 5 y vale 1 si el inmueble se construyó antes de 1950 y 5 si es posterior al año 2007).

Se chequearon los supuestos clásicos del modelo y, en relación a los residuos (Figura E.35, Apéndice E), hay observaciones con valores atípicos que evidencian una asimetría izquierda más exacerbada que en las ofertas (gráfico Q-Q). Esto puede deberse a que persisten valores bajos en algunas variables. Respecto al VIF, no se detectan problemas de multicolinealidad.

6.2.3. *B. Árbol de regresión - transacciones de apartamentos*

El árbol que resulta de validación cruzada tiene un parámetro $\alpha = 0{,}059$ y se ilustra en la Figura 22. Las variables que realizan las particiones son *superficie construida* a partir de 67 m^2 ($e^{4{,}21} \approx 67$) y *distancia a la playa* a partir de 1,64 kms. ($e^{0{,}497} \approx 1{,}64$).

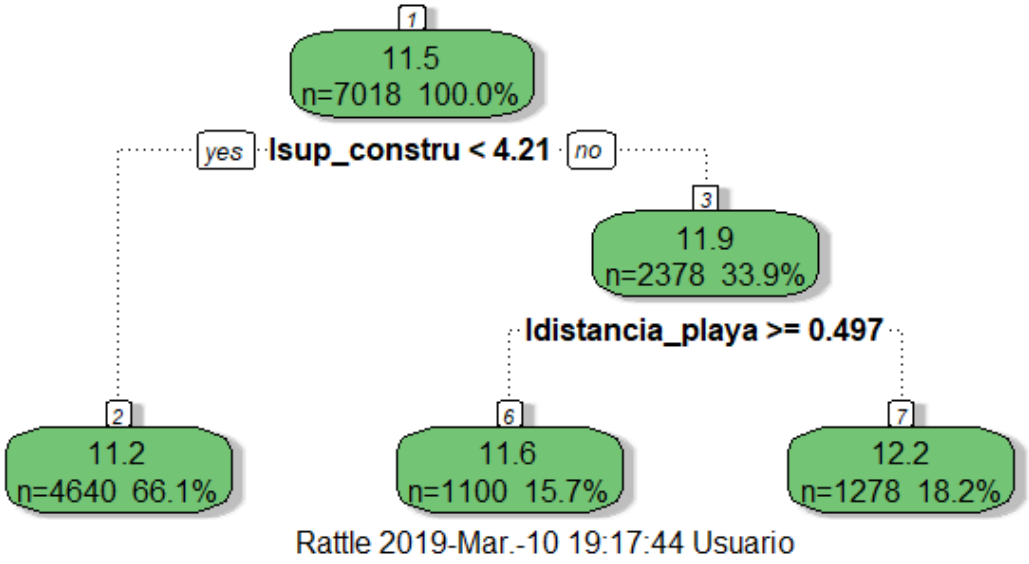


Figura 22: Árbol de regresión (transacciones de apartamentos)

6.2.4. C. Bosques Aleatorios - transacciones de apartamentos

La validación cruzada selecciona $p = 9$ para el modelo. El resultado es el siguiente:

Cuadro 6: Validación cruzada. Bosques aleatorios - transacciones de apartamentos

Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 5614, 5615, 5615, 5614,, ...
Resampling results across tuning parameters:

mtry	RMSE	Rsquared	MAE
2	0.5097196	0.5374978	0.3786599
9	0.4961172	0.5404552	0.3557796
16	0.5030150	0.5290020	0.3612274

6.2.5. D. Desarrollo del mejor modelo con transacciones de apartamentos

La performance de los tres modelos no alcanza las expectativas en la medida que en los tres casos el error es alto en términos absolutos y relativos. El modelo de bosques aleatorios es el mejor de los tres, superando en tres puntos porcentuales del EPAM, lo que representa una mejora de 6 % (47 % a 44 %). Si se considera el RECM la mejora es del orden de 10 %.

Cuadro 7: Performance de los modelos con transacciones de apartamentos

Modelo	RECM	EPAM
Lineal	59.030	47 %
Árbol	73.240	63 %
Bosque	**52.850**	**44 %**

Se desarrolla el modelo de bosques aleatorios agregando las siguientes 12 variables:

- *longitud*
- *latitud*
- *barrio*
- *municipio*
- *balcón*
- *planta*
- *zona legal*
- *amenities*
- *playa*
- *fecha (mes y año)*
- *valor total catastral*
- *log(dist. playa del este)*

Los resultados obtenidos muestran que el error se reduce en dos puntos porcentuales respecto al modelo de bosques aleatorio anterior. Sin embargo, aún persiste un error alto.

Cuadro 8: Performance de bosque desarrollado con transacciones de apartamentos

RECM	EPAM
50.631	42 %

La medida de importancia de las variables destaca el valor de catastro (*valor_total*), variable incorporada en este desarrollo del modelo. También, se destacan la variable antigüedad y las referidas a la ubicación del inmueble. En este caso se agregaron las variables latitud y longitud, que resultan relativamente importantes.

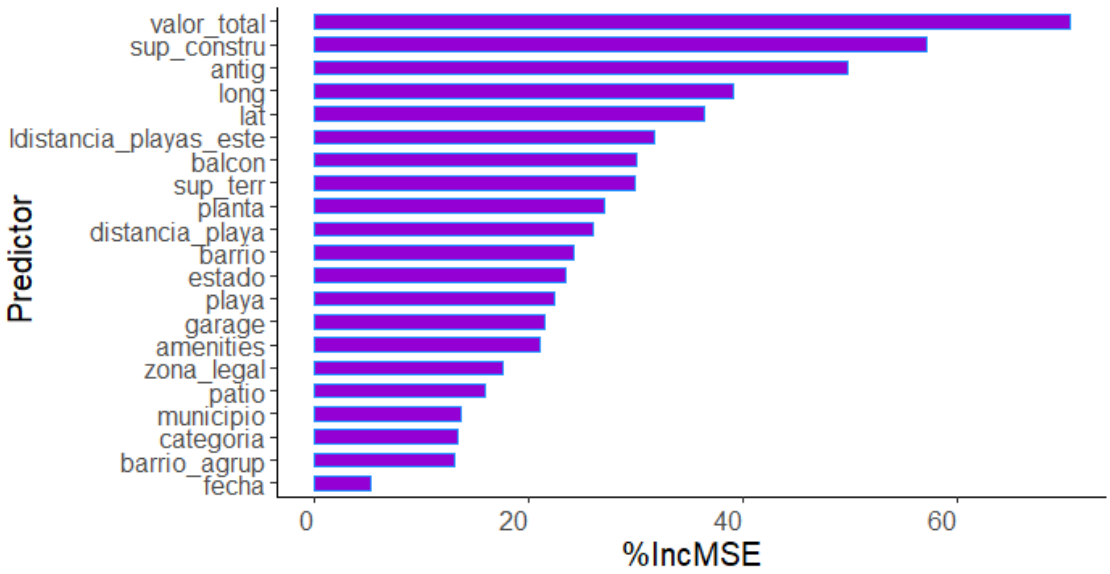


Figura 23: Importancia de las variables para predecir el precio de transacción de apartamentos - Modelo bosques aleatorios desarrollado

6.2.6. Casas - transacciones

La muestra de entrenamiento consta de 1.956 observaciones de un total de 2.792. El tamaño de la muestra representa un problema para la validación cruzada, dado que las particiones aleatorias pueden no tomar todos los niveles de las variables factores y las particiones pueden no tener suficiente variabilidad. Se opta por restringir el número de variables a *log(distancia a playa)*, *log(sup. construida)*, *antigüedad*, *garage* y *barrio costa* que indica si el inmueble se encuentra en un barrio de costa este o no. El modelo de bosques aleatorios mejora por poco al modelo lineal (ver desarrollo en el Apéndice E.2.2).

Cuadro 9: Performance de los modelos con transacciones de casas

Modelo	RECM	EPAM
Lineal	87.050	56 %
Árbol	99.570	67 %
Bosque	86.000	56 %
Bosque desarrollado	76.784	50 %

7. Resumen de resultados

Los modelos entrenados anteriormente muestran, en todos los casos, mejoras en la performance predictiva al implementar bosques aleatorios. Las mejoras relevantes se observan en las bases de datos de ofertas, donde se cuenta con un gran número de observaciones y un grupo confiable de variables. En el caso de las transacciones, tanto la cantidad de observaciones como la menor cantidad de características, pueden influir en los altos errores de predicción de todos los modelos. A su vez, esto puede estar relacionado a la baja mejora del modelo de bosques aleatorios respecto al lineal. En el Cuadro 10 se presentan los resultados resumidos de todos los modelos para cada base de datos.

Cuadro 10: Resumen de modelos y performance predictiva

Modelos con ofertas - **apartamentos**	RECM	EPAM
Lineal	USD 47.760	17 %
Árbol	USD 71.420	26 %
Bosque Aleatorio	USD 30.960	10 %
Bosque Aleatorio desarrollado	USD 29.530	9 %
Modelos con ofertas - **casas**		
Lineal	USD 75.240	22 %
Árbol	USD 126.740	29 %
Bosque Aleatorio	USD 56.800	14 %
Bosque Aleatorio desarrollado	USD 53.200	13 %
Modelos con transacciones - **apartamentos**	RECM	EPAM
Lineal	USD 59.030	47 %
Árbol	USD 73.240	63 %
Bosque Aleatorio	USD 52.854	44 %
Bosque Aleatorio desarrollado	USD 50.630	42 %
Modelos con transacciones - **casas**		
Lineal	USD 87.050	56 %
Árbol	USD 99.570	67 %
Bosque Aleatorio	USD 86.000	56 %
Bosque Aleatorio desarrollado	USD 76.780	50 %

8. Comentarios finales

En este trabajo se utilizaron modelos predictivos para precios de inmuebles de Montevideo, en un marco de aprendizaje estadístico. Éste se configura como complementario a los modelos que se enfocan en la estimación de efectos y determinantes. Asimismo, la metodología propuesta puede ser extendida a otros problemas de predicción. A su vez, se pone a disposición la base de datos de ofertas que puede ser actualizada.

Los resultados arrojan que el modelo de bosques aleatorios tiene una mejor performance predictiva respecto al modelo lineal hedónico, considerando los datos de Montevideo, tanto de ofertas como de transacciones. Esto parecería estar explicado por la naturaleza no lineal del problema de predicción. La superioridad del modelo de bosques aleatorios es aún mayor con los datos de ofertas. Con estos datos para apartamentos, el error porcentual absoluto medio (EPAM) evaluado en una muestra de testeo, se reduce en 40 % (de 17 % a 10 %) al utilizar el modelo de bosques aleatorios respecto al modelo lineal. Para las casas el EPAM se reduce 36 % (22 % a 14 %) respecto al modelo lineal. Las predicciones de los modelos ampliados, en el caso de las ofertas, presentan un error de USD 29.500 (raíz del error cuadrático medio, RECM) para apartamentos, y USD 56.800 (RECM) para las casas. En el caso de las transacciones, la performance del modelo de bosques aleatorios también mejora las predicciones respecto al lineal, aunque en proporciones modestas. Para los apartamentos mejora un 6 % (del EPAM) y para las casas la reducción es de 1 % (del RECM).

Los resultados obtenidos pueden establecer una base para desarrollos posteriores que exploren variantes en los modelos utilizados, por ejemplo ajustando los parámetros de variables y poda. Igualmente, se podría completar el análisis contemplando nuevas técnicas de aprendizaje estadístico. Adicionalmente es posible incorporar más variables hedónicas (incluyendo las espaciales y geográficas) y utilizar técnicas de econometría espacial. La variable distancia a la playa puede ser enriquecida y desarrollada con más variables de esta índole. Adicionalmente, los modelos podrían contemplar por un lado la dimensión temporal y por el otro, fundamentos macroeconómicos.

Por último, la brecha entre el precio pedido y el efectivamente llevado a cabo, puede resultar de interés la predicción conjunta y cruzada de precios (predicción de precio de transacción con precios de oferta). En relación a este análisis, dos posibles caminos son pertinentes. Por un lado, el estudio de la brecha de precios (oferta vs. transacción) en un contexto como el uruguayo (economía pequeña y dolarizada) y, por el otro, la utilización de índices de precios de oferta conjuntamente con los de transacción. Esto puede ser útil para predecir índices de precios en el tiempo, detectar desvíos de fundamentos y establecer una medida de tensión de mercado o indicador adelantado del sector y de la economía.

Referencias

Abadie, A., Kasy, M., 2017. The risk of machine learning. arXiv preprint arXiv:1703.10935.

Athey, S., 2018. The Impact of Machine Learning on Economics. In: The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, pp. 1–31.

Benítez-Silva, H., Eren, S., Heiland, F., Jiménez-Martín, S., 2015. How well do individuals predict the selling prices of their homes? Journal of Housing Economics 29, 12–25.

Berrutti Rampa, F., 2016. Subsidios a la oferta y decisiones de localización: El caso de la Ley de Vivienda de Interés Social. Serie documentos de investigación, Instituto de Economía, FCEA-UdelaR.

Breiman, L., 1996. Stacked regressions. Machine learning 24 (1), 49–64.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Routledge.

Carlomagno, G., Fernández, A., 2007. El precio de los bienes inmuebles: un estudio agregado y comparado para algunos barrios de Montevideo. mimeo, CINVE.

Carrillo, P. E., 2013. To sell or not to sell: Measuring the heat of the housing market. Real estate economics 41 (2), 310–346.

Čeh, M., Kilibarda, M., Lisec, A., Bajat, B., 2018. Estimating the performance of random forest vs multiple regression for predicting prices of apartments. International Journal of Geo-Information 7 (5), 168.

Cheng, J., Karambelkar, B., Xie, Y., 2018. Interactive Maps with JScript Leaflet. R package v. 2.0.2.

Chiarazzo, V., Caggiani, L., Marinelli, M., Ottomanelli, M., 2014. A Neural Network based model for real estate price estimation considering environment. Transportation Research Procedia 3, 810–817.

De Bruyne, K., Van Hove, J., 2013. Explaining the spatial variation in housing prices: an economic geography approach. Applied Economics 45 (13), 1673–1689.

De Rosa, M., Siniscalchi, S., Vigorito, A., Willebald, H., 2016. El estado del arte de los estudios distributivos en Uruguay. Instituto de Economía, FCEA-UdelaR y CEF.

Domínguez, M., Fornasari, N., Lanzilotta, B., Pareschi, F., 2016. Determinantes macroeconómicos del precio de la vivienda en Montevideo. Tech. rep., CEEIC.

Drenik, A., Pérez, D. J., 2017. Pricing in Multiple Currencies in Domestic Markets. Working paper.

Dubin, R. A., 1998. Predicting house prices using multiple listings data. The Journal of Real Estate Finance and Economics 17 (1), 35–59.

Fan, G.-Z., Ong, S. E., Koh, H. C., 2006. Determinants of house price: A decision tree approach. Urban Studies 43 (12), 2301–2315.

Fischer, S., 2017. Housing and Financial Stability: a speech at the DNB-Riksbank Macroprudential Conference Series, Amsterdam, Netherlands.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Vol. 1. Springer series in statistics New York.

García López, R., 2018. Consecuencias de una política de exoneración impositiva sobre la localización y el acceso a la vivienda inclusiva. Documento de trabajo wp18rg1sp, Lincoln Institute of Land Policy.

Genuer, R., Poggi, J.-M., Jan. 2017. Arbres CART et Forêts aléatoires, Importance et sélection de variables. arXiv preprint arXiv:1610.08203.

González-Pampillón, N., 2017. Spillover effects from a place-based housing subsidy. Working paper, Institut d'Economia de Barcelona (IEB) & Universitat de Barcelona.

Goyeneche, J. J., Moreno, L., Scavino, M., 2017. Predicción del valor de un inmueble mediante técnicas agregativas. Serie DT IESTA (17/1).

Griliches, Z., 1961. Hedonic price indexes for automobiles: An econometric of quality change. In: The price statistics of the federal goverment. NBER, pp. 173–196.

Han, L., Strange, W. C., 2016. What is the role of the asking price for a house? Journal of Urban Economics

93, 115–130.
Haurin, D., McGreal, S., Adair, A., Brown, L., Webb, J. R., 2013. List price and sales prices of residential properties during booms and busts. Journal of Housing Economics 22 (1), 1–10.
Herath, S., Maier, G., 2010. The hedonic price method in real estate and housing market research: a review of the literature. SRE - Discussion Papers, 2010/03. WU Vienna University of Economics and Business.
Horowitz, J. L., 1992. The role of the list price in housing markets: theory and an econometric model. Journal of Applied Econometrics 7 (2), 115–129.
IMM, 2013. Informe Censos 2011: Montevideo y Área Metropolitana. Unidad de estadística, IM.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Vol. 1. New York: Springer.
Kiel, K., Zabel, J., 2004. Location, location, location: The 3L Approach to house price determination. Journal of Housing Economics 17 (2), 175–190.
Kuhn, M., 2018. The caret package manual. R Foundation, Vienna.R package v. 5.2.2.
Kuhn, M., otros., 2018. Caret: Classification and Regression Training. R package v. 6.0-84.
Landaberry, V., Tubio, M., 2015. Estimación de índice de precios de inmuebles en Uruguay. Documento de trabajo del Banco Central del Uruguay.
Lanzilotta, B., Veneri, F., 2016. Variación geográfica del precio de la vivienda en Montevideo: análisis de determinantes y medición de efectos barrios. Estudio aplicado entre 2001-2014. Tech. rep., CEEIC.
Leamer, E., September 2007. Housing IS the Business Cycle. Working Paper 13428, NBER.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2 (3), 18–22.
Licandro, G., Ponce, J. (Eds.), octubre 2015. Precios de activos internos, fundamentos globales y estabilidad financiera. No. 4 in Investigación Conjunta-Joint Research. CEMLA.
Mooya, M. M., 2016. Standard Theory of Real Estate Market Value: Concepts and Problems. In: Real Estate Valuation Theory. Springer, pp. 1–21.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives 31 (2), 87–106.
Ooms, J., 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO].
Park, B., Bae, J. K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications 42 (6), 2928–2934.
Ponce, J., Tubio, M., 2013. Precios de inmuebles: aproximaciones metodológicas y aplicación empírica. Documento de trabajo del Banco Central del Uruguay.
Richardson, L., 2007. Beautiful soup documentation. April.
Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy 82 (1), 34–55.
Selim, H., 2009. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications 36 (2), 2843–2852.
Therneau, T., Atkinson, B., 2018. rpart: Recursive Partitioning and Regression Trees. R package v. 4.1-15.
Varian, H. R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives 28 (2), 3–28.
Wang, X., Wen, J., Zhang, Y., Wang, Y., 2014. Real estate price forecasting based on SVM optimized by PSO. Optik-International Journal for Light and Electron Optics 125 (3), 1439–1443.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wickham, H., 2018. httr: Tools for Working with URLs and HTTP. R package v. 1.4.0.
Wickham, H., 2019. rvest: Easily Harvest (Scrape) Web Pages. R package v. 0.3.3.
Zhu, M., 2014. Housing markets, financial stability and the economy, housing markets and the macroeconomy: challenges for monetary policy and financial stability conference, Eltvile, Bundesbank.

Apéndice A. Descripción de variables y criterios de limpieza

Apéndice A.1. Variables de ofertas (selección)

Variable	Descripción	Tipo de dato	Criterios considerados en la limpieza
año,mes	año, mes de bajada	construida	Año y mes en que se realizó la bajada de datos
barrio_agrup	barrios agrupados por cercanía	construida	Barrios que tienen al menos 3000 obs.. Se agrupan barrios con menos obs. en base a cercanía, de forma de llegar al menos a 3000 obs. y lograr 12 grupos de barrios. La Teja y Cerro se agrupa con Peñarol, Sayago, Colón y Piedras Blancas. El criterio aquí es la lejanía al centro de la ciudad.
distancia_playa	mínima distancia a una playa montevideana	construida	Mínima distancia euclidiana entre el punto en que se encuentra el inmueble (*centroide*) y el punto que identifica a una playa de Montevideo. Se imputa la media de la distancia por barrio a los valores faltantes (7000 obs.)
distancia_playa_este	mínima distancia a una playa del este de Montevideo	construida	ídem anterior con playa del Este de Montevideo (a partir de la Playa Ramírez). Se imputa la media de la distancia por barrio a los valores faltantes finales (aprox. 7000 observaciones)
playa	nombre de playa con la mínima distancia	construida	Las imputaciones a las valores NA son coherentes con el cambio realizado a la variable distancia
playa_este	nombre playa del este	construida	Las imputaciones a NA son coherentes con el cambio a la var. distancia
preciom2	precio / sup_constru	construida	Se toma el 98 % central de los valores luego limpiar precio y sup_constru
lat	latitud	sitio web deduce de la dirección	Se toman los valores coherentes con los límites de Montevideo y alrededores, se imputa NA a los que no corresponden.
long	longitud	sitio web la deduce de la dirección	Se toman los valores coherentes con los límites de Montevideo y alrededores, se imputa NA a los que no corresponden.
id	id de la publicación	generada por el sitio	Se procuró descartar los duplicados en id y precio (si se encuentra duplicado el mismo id es porque cambió su precio y se mantiene la observación).
start_time	fecha de creación	generada por sitio	Variabe disponible a partir de bajada de datos de abril
stop_time	fecha de fin esperada	generada por sitio	Variable disponible a partir de bajada de datos de abril
banos	número de baños	oblig. numérica	Se redujo a valores entre 1 y 3. Los valores mayores a 3 se asignan a 3
direccion	dirección del inmueble	obligatoria, con chequeo	Dirección incluye calle y número. Se chequea para muchos casos que se repite la misma dirección en diferentes publicaciones. En los casos que se suponga que la dirección no corresponde, se imputa NA en long y lat.
titulo	título de publicación	obligatoria, con chequeo de palabras	No se realizan cambios a esta variable. Se utiliza para chequear coherencia y eliminar duplicados. En los casos que el título y precio se repetía, se hizo un chequeo para eliminar en caso de tratarse de duplicados.
dormitorios	número dormitorios	obligatoria, cualquier formato	Se transformaron valores como “monoambiente”, “no tiene”, etc. a 0; “más de 4”, a 5. Se redujo a valores entre 0 (monoambientes) y 3, y se toma como factor. Los mayores a 3 se asignan al valor 3. Pocos valores que quedan con NA, se imputan a 0 y 1 luego de chequear coherencia (cruzado con sup_constru y precio). En el caso de las casas se transforma en un factor con dos opciones: 1 para los casos de dos dormitorios o menos y 2 para más de dos.
barrio	barrio indicado del inmueble	obligatoria, lista desplegable para seleccionar	Barrios que tienen al menos 400 observaciones. Se agrupan los barrios con menos observaciones si son cercanos, de forma que lleguen a las 400 observaciones.
condicion	condición del inmueble (nuevo/usado)	obligatoria, desplegable (nuevo, usado, sin especificar)	Se hizo una asignación de los inmuebles que indicaban “sin especificar”. Si la variable antigüedad era menor a 1 año, se asigna como nuevo, de lo contrario se asigna a “usado”. La mayoría de “sin especificar”son usados.
cocheras	número de cocheras	obligatoria, numérica	Número de cocheras que tiene el inmueble. Se redujo a valores entre 0 (no tiene) a 2. Los valores mayores a 2 se agrupan en el valor 2.
precio	precio solicitado en USD	obligatoria, numérica	Se toma el 96 % de los valores centrales. Se eliminan observaciones con precios como sucesiones de números (“12345”, etc.) o números con dígitos repetidos (“1111”, ‘9999”, etc.). Se toman solo precios en USD.
sup_constru	superficie en m2	obligatoria, numérica	Se imputa NA a valores de 3 cifras y más que repiten dígitos y son mayores a 300 (“333”,“444”, “999”, “1111”, etc.). Se limpian valores que repiten dígitos menores a “300”(“222”, “111”) luego de verificar. Se toma las obs. del 98 % central de los valores. Se imputa NA a valores menores a 9m2 y mayores a 2000m2. Se cruza con sup_total, luego de su limpieza para completar NA. En base final se eliminan obs. con NA en esta variable.
tipo inmueble	tipo de inmueble	obligatoria, desplegable	Se seleccionan solamente apartamentos y casas

Variable	Descripción	Tipo de dato	Criterios considerados en la limpieza
sup_tot	superficie total en m2	obligatoria, numérica	Se imputan NA a valores menores a 9 y mayores a 20000. Se imputa NA igual que en sup_constru. Se cruza la variable de forma que la superficie construida sea menor que la total.
pisos	pisos del edificio	obligatoria numérica	Se transformó de forma que se restrinja a valores entre 0 y 20. En casos que se indicaban valores de 3 o 4 cifras, se deja solo el primer dígito o primeros dos dígitos (ejemplo. 902 se transforma a 9, 1002 a 10).
orientacion	orientación del inmueble	opcional, desplegable	No se realizan cambios a esta variable. Se utiliza para chequear coherencia con características
tipo_edif	tipo de edificio	opcional, desplegable	Los valores NA, fueron imputados a "No indica"
ambientes	Número de ambientes	opcional, cualquier formato	Se restringe la variable al intervalo 1-5. Mayores a 5 se incluyen en 5.
antigüe-dad	años desde que se construyó el inmueble hasta 2018 (o año en que se construyó)	opcional, cualquier formato	Se transformaron los diferentes valores para que indiquen los años desde la construcción en un intervalo [-2 , 120]. En los casos que indicaba el año de construcción (1960 por ejemplo) se hace la diferencia con el año 2018 (2018 - 1960 = 58). En algunos casos el valor es negativo (en general, se refiere a inmuebles en construcción). Valor .ª estrenar.º "nuevo"se transforma al valor 0. Si se indica "más de.º similar se deja el año que se indica. Lo mismo cuando se indica "máx. X años". Se imputa NA a valores no coherentes ("111", "999", etc.).
ap_piso	apartamentos por piso en el edificio	opcional, cualquier formato	En el caso que se indicara el número de unidad (tres o cuatro cifras), se imputa el último dígito de la cifra
sup_ balcon	superficie del balcón en m2	opcional, cualquier formato	Se chequea que sea numérico y se transforman valores no coherentes. Se imputa NA a valores mayores a 199m2
expensas	gastos comunes del apartamento / PH	opcional, cualquier formato (a partir de octubre/noviembre de 2018 chequea que sea numérico)	Se transforman los valores "no tiene", "no hay", "sin gastos", etc. al valor 0. En caso que indique un intervalo de valores, se toma el menor valor. En caso que diga "bajos gastos comunes"se imputa el valor $ 1000. En casos en que el monto es una sucesión de números ("123", "1234", etc.) o tres cifras o más repetidas ("333", "5555", etc) se imputa a NA. En los casos que se indica el monto en USD, se multiplica por $32. Se imputan NA en casos que indican "a consultar.º no indican monto. Se imputa NA en valores mayores a $100.000. Valores mayores a 50000, se divide entre 10. Se imputa la mediana considerando barrio y si tiene ascensor o no a los NA.
ascensores	número de ascensores	opcional, cualquier formato. Luego de ene-2018, Sí/No.	variable válida en apartamentos. Se redujo a valores entre 0 y 2. Más de 2, en 2. Se imputó la mediana por barrio y ascensor (si tiene o no) a los NA.
aircond	tiene aire acondicionado? Sí/No	opcional, se selecciona en caso de Sí	Los valores NA, fueron imputados a "No"
asc_serv	tiene ascensor de servicio? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
balcon	tiene balcón? Sí/No	ídem anterior	No se realizan cambios relevantes
bano_soc	tiene baño social? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
calefacc	tiene calefacción? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
comedor	tiene comedor? Sí/No	ídem anterior	No se imputan valores
lavadero	tiene lavadero? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
parrillero	tiene parrillero? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
patio	tiene patio? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
salon_ comunal	tiene salón comunal? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
seguridad	tiene seguridad?Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
toilette	tiene toilette? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
living	tiene living? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
dormisuite	tiene un dormitorio en suite? Sí / No?	ídem anterior	Los valores NA, fueron imputados a "No"
placard	tiene placards? Sí/No	ídem anterior	Los valores NA, fueron imputados a "No"
garage	tiene garage? Sí/No	ídem anterior (luego ene/2018 no está)	Se cruza con la variable cocheras (si cochera es mayor a uno, tiene garage). NA se imputa como 'No".
amoblado	tiene amoblamiento incluido? Sí/No	ídem anterior, posibilidad de aclarar (luego ene/2018 no está)	Se transformó para que todos los tipos de amoblamiento indicados sean "Si", incluso cuando se indicaba "opcional". NA se imputa "No"
piscina	tiene piscina? Sí/No	ídem anterior. Luego ene-2018, Sí/No.	Se transformó para que todos los tipos indicados sean "Si". Los valores NA fueron imputados como "No"

Apéndice A.2. Variables de transacciones

Variable	Descripción	Tipo de dato	Criterios considerados en la limpieza
padrón	número de padrón	Variable numérica de DGR - DNC	Variable que identifica el inmueble, no se modifica.
unidad	número de unidad en caso que corresponda	variable numérica de DGR - DNC	Variable que identifica el inmueble para los casos que corresponde (propiedad horizontal), no se modifica.
block	número de black en caso que corresponda	variable alfabética de DGR - DNC	Variable que identifica el inmueble para los casos que corresponde, no se modifica.
precio	precio solicitado	MONTO, obligatoria en decl. a DGR	Se eliminan los extremos: 3 % de la izquierda y 2 % de la derecha. Se eliminan más datos por izquierda por la gran cantidad de bajos valores.
precio_cut	categoría de precios	construida	Rangos de precios a partir de quintiles de precios de cada tipo de propiedad (PH/apartamento y CO/casas por separado)
sup_ construu	sup. construida en m2	variable de DNC	Se eliminan observaciones con menos de 15 m2 y más de 2000 m2.
sup_terr	sup. del terreno en m2	variable de DNC	Se eliminan observaciones con más de 10000 m2.
preciom2	precio/sup_construu	construida	Se toman las observaciones con el 98 % central de los valores
barrio	barrios	DNC, definida como área diferencial	Barrios que tienen al menos 90 observaciones. Se agrupan los barrios con menos obs. si son cercanos, de forma que lleguen a las 90 obs.
barrio_ agrup	barrios agrupados	contruida	Se agrupan barrios para tener al menos 800 observaciones en cada factor. Los barrios con menos obs. se agrupan en base a cercanía. La agregación no corresponde a barrios homogéneos.
estado	estado de conservación del inmueble	variable de DNC	Se restringe la variable a valores entre 1 a 3. Donde 1 indica estado excelente y bueno y 3 mal estado. Valores NA (7 %) se imputan a 2. Originalmente la variable contaba con 10 valores.
antig	antigüedad	variable de DNC	Transformada a categórica entre que vale 1(anterior a 1950), 2 (1950-1970), 3 (1971 - 1990), 4 (1991-2007) y 5 (luego de 2007). Se toma la transformación hecha en Landaberry y Tubio (2015)
distancia_ playa	ídem ofertas	construida	ídem ofertas.
playa	ídem ofertas	construida	ídem ofertas
distancia_ playa_este	ídem ofertas	construida	ídem ofertas
playa_este	ídem ofertas	construida	ídem ofertas
tipo_ inmueble	tipo de inmueble	obligatoria	Se cambia el nombre de propiedad horizontal a apartamentos y propiedad común a casas
lat	latitud	construida con centroide del padrón	Se calcula el centroide del polígono que conforma el padrón. El archivo utilizado para ello es el archivo Shapes del parcelario rural y urbano (MVOTMA-IM).
long	longuitud	ídem latitud	ídem latitud
garage	tiene garage? Sí / No	variable de DNC	Se imputan los valores NA, a 0 (No tiene)
balcon	tiene balcón? Sí / No	variable de DNC	Se imputa 0 (No tiene) a los NA.
patio	tiene patio? Sí / No	variable de DNC	Se imputa 0 (No tiene) a los NA. Incluye jardín.
ph	Es propiedad horizontal? Sí / No (1 o 0)	variable de DNC	No se realizan cambios
ammenities	tiene piscina, barbacoa o similar? Sí / No	variable de DNC	Se agrupan la amenities y se imputa 0 (No tiene) a los NA
categoría	categoría de construcción	variable de DNC	Se tranforma para que quede con valores de 1 (confortable y muy confortable), 2 (común, e imputación de NAs) y 3 (económica y muy económica). Originalmente la variable contaba con 10 categorías.
valor_total	valor total de catastro en pesos	variable de DNC	No se realizan cambios ni modificaciones. El valor refiere a la última actualización disponible.
valor_terr	valor del terreno en pesos	variable de DNC	No se realizan cambios ni modificaciones. El valor refiere a la última actualización disponible.
municipio	municipio	variable de DNC	No se realizan cambios ni modificaciones.
zona_legal	centro comunal zonal	variable de DNC	No se realizan cambios ni modificaciones.
unidad	núm. unidad de PH	variable de DNC	No se realizan cambios ni modificaciones.
planta	planta en donde se encuentra el inmueble	construida en base a DNC	Toma valores entre 0 a 5. 0 es planta baja, 1 primer piso, 2 segundo piso, 3 tercer piso, 4 cuarto a sexto piso y 5 corresponde a los restantes pisos. Se imputó la mediana por barrio y antigüedad (mayor a 3 y menor a 3) en los casos de PH/Apartamento. En CO que son NA, se imputa 0.

Apéndice B. Visualizaciones descriptivas adicionales

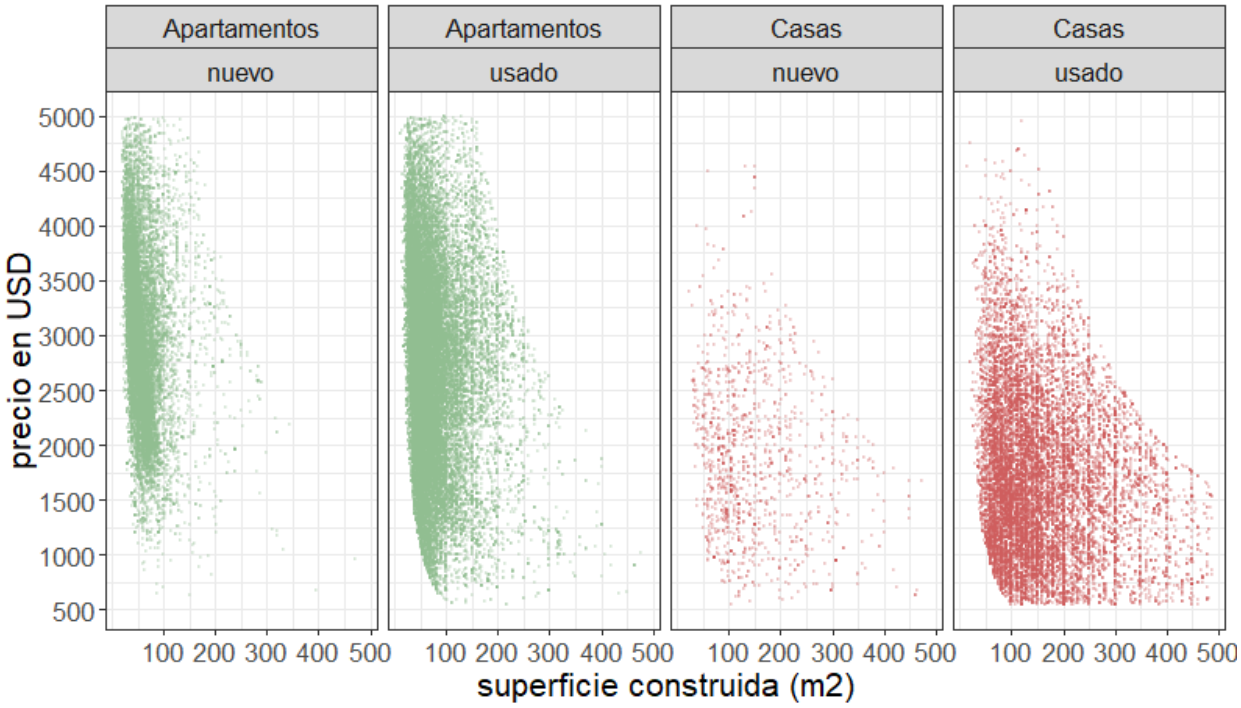


Figura B.24: Precio del m^2 ofertado y superficie construida (m^2)
Fuente: Elaboración propia en base a mercadolibre.com.uy.

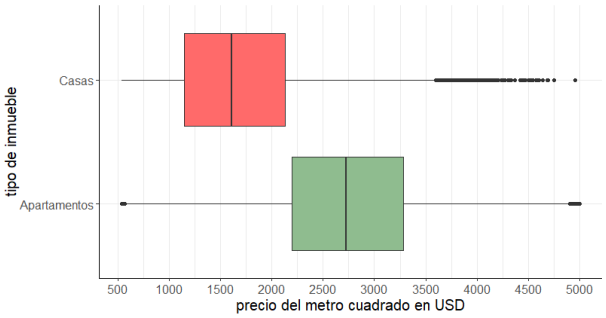


Figura B.25: Distribución de la variable precio del m^2 por tipo de inmueble ofertado
Fuente: Elaboración propia en base a mercadolibre.com.uy.

Figura B.26: Distribución del precio del m^2 transado según tipo de inmueble
Fuente: Elaboración propia en base a DGR, DNC propia

Apéndice C. Mapas adicionales

Datos de oferta - precio del m2

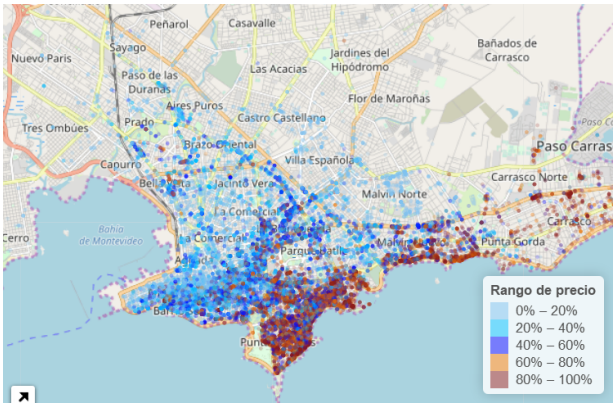


Figura C.27: Apartamentos y precio m^2

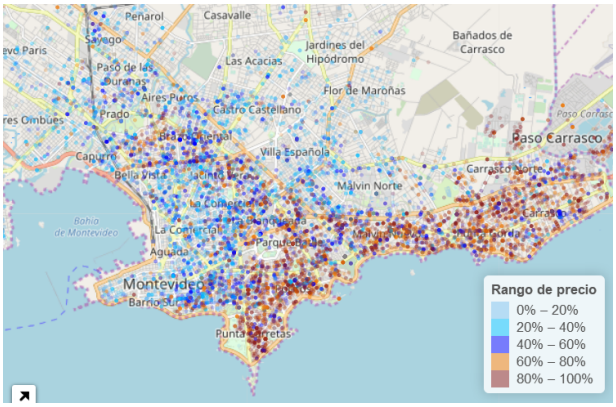


Figura C.28: Casas y precio m^2

Fuente: Elaboración propia en base a mercadolibre.com.uy

Datos de transacciones - precio del m2

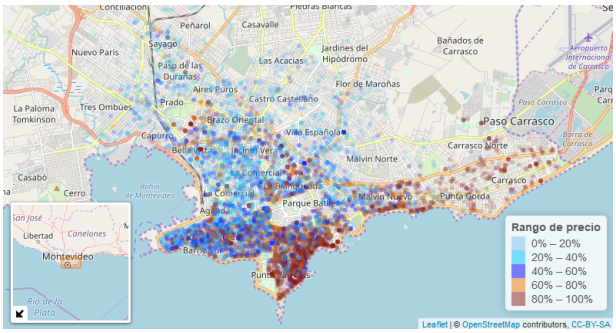


Figura C.29: Apartamentos y precio m^2

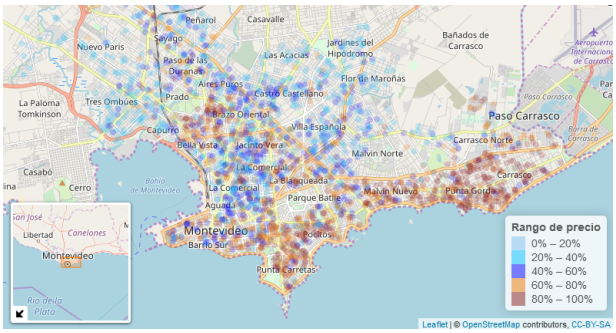


Figura C.30: Casas y precio m^2

Fuente: Elaboración propia en base a DGR, DNC, SIG-IM

Apéndice D. Cálculo de la variable distancia

1. Mínima distancia euclidiana (o a *vuelo de pájaro*) entre la vivienda y las playas del Este de Montevideo (a partir de la playa Ramírez)
2. Mínima distancia euclidiana entre la vivienda y todas las playas de Montevideo, incluye las playas contempladas en la variable anterior y se suman las del oeste de Montevideo (Playa del Cerro, Pajas Blancas, etc.).

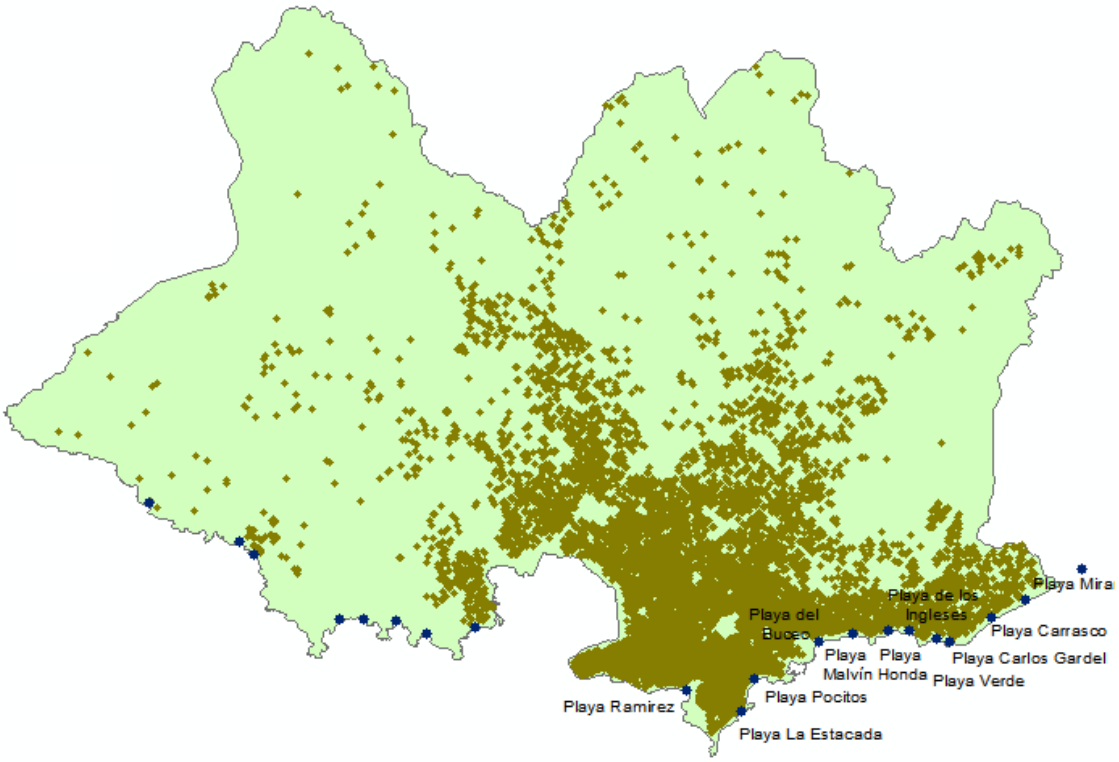


Figura D.31: Ubicación de transacciones y playas de Montevideo
Las playas se indican con un punto azul. Se nombran solo las playas del este.
Fuente: Elaboración propia, en base a DGR, DNC, MVOTMA, SIG-IM.

- Ubicación de playas: MVOTMA (www.dinama.gub.uy/geoservicios)
- Programa utilizado para la medición: ArcGIS (*Generar tabla de cercanía*)

Apéndice E. Salidas adicionales

Apéndice E.1. Modelos de precios de oferta

Apéndice E.1.1. Apartamentos - ofertas

Cuadro E.11: Modelo lineal hedónico para el precio de ofertas de apartamentos

	Variable dependiente: ***log(precio)***	
	coeficientes	(std. error)
log(distancia a la playa) (en km)	-0.060^{***}	(0.002)
Cordón (base es Pocitos)	-0.197^{***}	(0.004)
Carrasco, PGorda, Malvin	0.023^{***}	(0.004)
Punta Carretas	0.113^{***}	(0.004)
LaComer., TCruces, Jac.V, LaBlanq	-0.187^{***}	(0.005)
BrazO., Atahua., Prado	-0.183^{***}	(0.007)
Buceo, PBatlle	-0.090^{***}	(0.005)
Pque. Rodo, Palermo	-0.087^{***}	(0.005)
Aguada, Reducto, PasoM., Belv.	-0.351^{***}	(0.007)
Centro, Ciudad Vieja	-0.209^{***}	(0.005)
MalNor, Maroñ, Uni., VilEsp., Cerri.	-0.576^{***}	(0.007)
Cerr., LaTe., Col., Say., Peñ., PaBla.	-0.699^{***}	(0.015)
condición usado (base es nuevo)	-0.079^{***}	(0.002)
1 dormitorio (base es 0)	0.060^{***}	(0.004)
2 dormitorios	0.079^{***}	(0.004)
3 o más dormitorios	0.021^{***}	(0.005)
baños (cantidad)	0.151^{***}	(0.003)
log(superficie construida) en (m^2)	0.615^{***}	(0.004)
garage Sí (base es No)	0.110^{***}	(0.002)
ascensores (cantidad)	0.012^{***}	(0.002)
constante	9.344^{***}	(0.012)
Observaciones	49,574	
R^2	0.800	
Adjusted R^2	0.800	
Residual Std. Error	0.216 (df = 49553)	
F Statistic	$9{,}907.693^{***}$ (df = 20; 49553)	
Nota:	*p<0.1; **p<0.05; ***p<0.01	

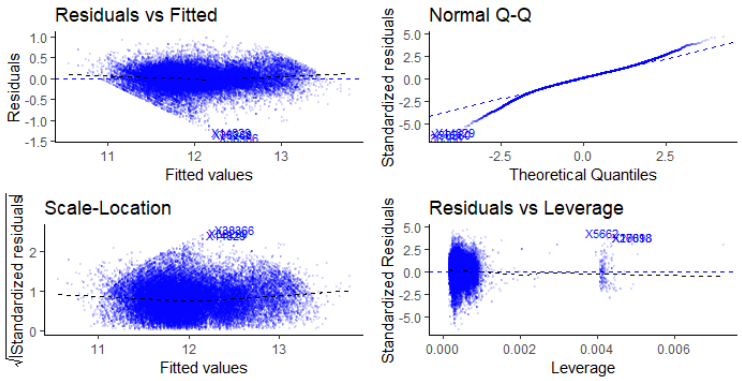


Figura E.32: Gráficos de residuos del modelo lineal con ofertas de apartamentos

Apéndice E.1.2. Casas - ofertas

Cuadro E.12: Modelo lineal Hedónico para el logaritmo del precio de ofertas de casas

	Variable dependiente: ***log(precio)***	
	coeficientes	(std. error)
log(distancia playa)	−0.028***	(0.005)
Cordón (base es Pocitos)	−0.369***	(0.015)
Carrasco, PGorda, Malvin	0.028***	(0.010)
Punta Carretas	0.159***	(0.018)
LaComer., TCruces, Jac.V, LaBlanq	−0.336***	(0.012)
BrazO., Atahua., Prado	−0.291***	(0.014)
Buceo, PBatlle	−0.133***	(0.012)
Pque. Rodo, Palermo	−0.280***	(0.015)
Aguada, Reducto, PasoM., Belv.	−0.550***	(0.015)
Centro, Ciudad Vieja	−0.469***	(0.018)
MalNor, Maroñ, Uni., VilEsp., Cerri.	−0.642***	(0.013)
Cerr., LaTe., Col., Say., Peñ., PaBla.	−0.719***	(0.014)
usado (base nuevo)	−0.050***	(0.009)
más de 2 dormitorios (base menos de 2)	0.031***	(0.006)
baños (cantidad)	0.114***	(0.004)
log(superficie construida)	0.555***	(0.006)
garage Sí (base No)	0.079***	(0.005)
constante	9.625***	(0.026)
Observaciones	15,412	
R^2	0.813	
Adjusted R^2	0.813	
Residual Std. Error	0.271 (df = 15394)	
F Statistic	3,938.160*** (df = 17; 15394)	
Nota:	*p<0.1; **p<0.05; ***p<0.01	

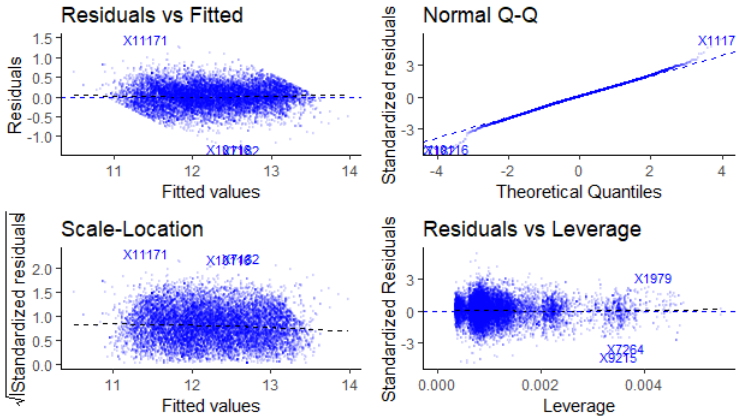


Figura E.33: Gráficos de los residuos del modelo lineal con ofertas de casas

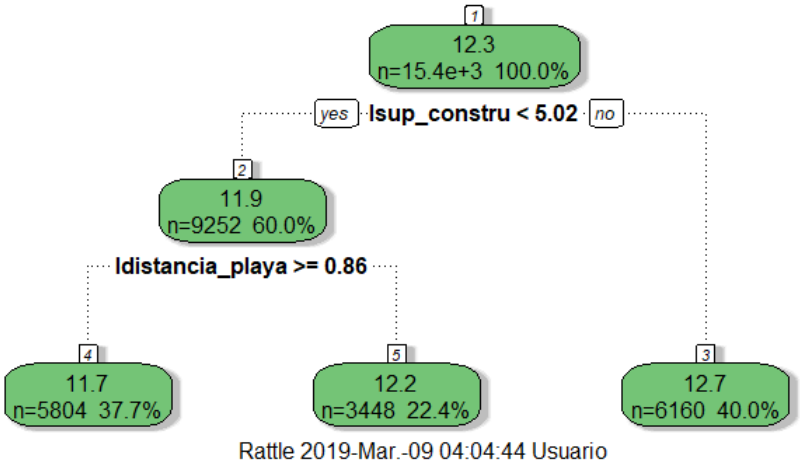


Figura E.34: Árbol de regresión (ofertas de casas)
El árbol seleccionado por la validación cruzada tiene un parámetro $\alpha = 0{,}065$.

Apéndice E.2. Modelos de precios de transacción

Apéndice E.2.1. Apartamentos - transacciones

Cuadro E.13: Modelo lineal hedónico para el precio de transacciones de apartamentos

	Variable dependiente: ***log(precio)***	
	coeficientes	(std. error)
log(distancia a la playa)	-0.101^{***}	(0.014)
BSur-PRodo-Paler. (base Pocitos-PCarr-Buceo)	-0.146^{***}	(0.027)
Carrasco, PGorda, Malvin	0.008	(0.025)
Centro, CiudadV.	-0.184^{***}	(0.021)
Jac.V, LaBlanq., LaComer., VMuñ.	-0.199^{***}	(0.027)
Maroñ, MalvNo., Cerri., Uni., etc.	-0.515^{***}	(0.036)
Prado, Capu., Say., Peñ., PMol., Cerro, etc.	-0.311^{***}	(0.034)
estado 2 (base 1)	-0.113^{***}	(0.021)
estado 3	-0.072^{***}	(0.022)
categoría 2 (base 1)	-0.126^{***}	(0.024)
categoría 3	-0.211^{***}	(0.031)
log(sup. construida)	0.681^{***}	(0.013)
antigüedad	0.117^{***}	(0.006)
garage Sí (base No)	0.168^{***}	(0.016)
patio Sí (base No)	0.086^{***}	(0.013)
log(sup. terreno)	0.014^{*}	(0.008)
constante	8.570^{***}	(0.083)
Observaciones	7,018	
R^2	0.525	
Adjusted R^2	0.524	
Residual Std. Error	0.503 (df = 7001)	
F Statistic	484.549^{***} (df = 16; 7001)	
Nota:	*p<0.1; **p<0.05; ***p<0.01	

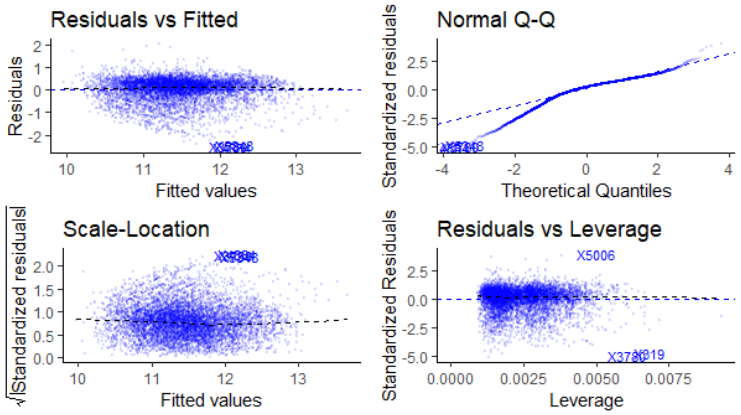


Figura E.35: Gráficos de residuos, modelo lineal con transacciones de apartamentos

Apéndice E.2.2. Casas - transacciones

Cuadro E.14: Modelo lineal hedónico para el precio de transacciones de casas

	Variable dependiente: ***log(precio)***	
	coeficientes	(std. error)
log(distancia a la playa)	−0.191***	(0.027)
barrio costero Sí	0.421***	(0.051)
log(sup. construida)	0.624***	(0.022)
antigüedad	0.096***	(0.012)
garage	0.166***	(0.031)
constante	8.323***	(0.120)
Observaciones	1,956	
R^2	0.574	
Adjusted R^2	0.571	
Residual Std. Error	0.577 (df = 1941)	
F Statistic	186.488*** (df = 14; 1941)	
Note:	*p<0.1; **p<0.05; ***p<0.01	