Título de la revista
XX(X):1-12

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Un conjunto de microdatos georreferenciados de anuncios inmobiliarios de las tres mayores ciudades españolas

Resumen

Este artículo presenta un producto de datos abiertos con grandes conjuntos de microdatos georreferenciados de anuncios inmobiliarios de 2018 en España. Estos datos se publicaron originalmente en el sitio web inmobiliario idealista.com. Las observaciones se obtuvieron para las tres ciudades más grandes de España: Madrid (n = 94 815 observaciones), Barcelona (n = 61 486 observaciones) y Valencia (n = 33 622 observaciones). Los conjuntos de datos incluyen las coordenadas de los inmuebles (latitud y longitud), los precios de venta de cada vivienda listada y diversas variables de características interiores. Los listados se enriquecieron con información oficial del catastro español (por ejemplo, calidad de los materiales de construcción) y otras características geográficas relevantes, como la distancia a puntos de interés urbano. Junto con los listados inmobiliarios, el producto de datos incluye también los límites de los barrios de cada ciudad. El producto de datos se ofrece como un paquete `R` totalmente documentado y está disponible para fines científicos y educativos, en particular para estudios geoespaciales.

Palabras clave

Precios de la vivienda; análisis hedónico de precios; idealista.com; datos georreferenciados; datos a nivel de punto; datos abiertos; España

Introducción

El interés por las características del mercado inmobiliario y los precios de la vivienda ha sido un área de investigación creciente en las últimas décadas, generando una gran cantidad de literatura teórica y empírica. La inclusión del componente espacial para analizar el mercado inmobiliario y la incorporación de variables geográficas ha mejorado significativamente la comprensión

de este mercado: para ello, es fundamental disponer de información/datos a nivel de punto. Por ello, cada vez es más habitual que en el análisis espacial de entornos urbanos se utilicen microdatos, georreferenciados como puntos (Lo'pez et al., 2015). En algunos casos, los investigadores han tenido que recurrir a procesos de webscraping para obtener datos para la investigación (por ejemplo, Li et al., 2019; Lo'pez et al., 2015). Lamentablemente, el webscraping es un proceso precario propenso a errores de descarga, datos que faltan, registros duplicados, etc. Además, los investigadores no siempre comparten sus conjuntos de datos webscrapped, lo que limita la reproducibilidad de su investigación. A medida que asistimos a un creciente interés por la apertura y la reproducibilidad en la ciencia de datos geográficos (Arribas- Bel et al., 2021a; Pa'ez, 2021; Brunsdon y Comber, 2021), se hace cada vez más urgente disponer de productos de datos abiertos para apoyar la investigación (Arribas-Bel et al., 2021b).

Algunos investigadores ya han respondido a esta necesidad de disponer de conjuntos de datos georreferenciados de acceso público para apoyar una investigación abierta y reproducible. Ya se dispone de algunos conjuntos de datos de este tipo para apoyar el análisis de los mercados inmobiliarios, pero a veces están georreferenciados a nivel de grandes zonas geográficas, como Fuerst y Haddad (2020), un conjunto de datos abierto que incluye $n = 4,201$ precios inmobiliarios geocodificados a nivel de nueve regiones de Inglaterra y Gales. Otros conjuntos de datos están geocodificados como puntos, incluidos Bonifaci y Copiello (2015), que incluye $n = 1,042$ observaciones para Padua, en Italia; Del Giudice et al. (2018), que comparten un conjunto de datos con $n = 576$ observaciones relativas a los precios de alquiler en Nápoles, Italia; y Solano Sa'nchez et al. (2019) presentan un conjunto de datos con $n = 1,623$ precios de alquiler diarios en Sevilla, España.

Para contribuir al creciente inventario de conjuntos de microdatos internacionales de mercados inmobiliarios, este documento presenta un producto de datos abiertos con listados de viviendas georreferenciadas, denominado *{idealista18}*. Los datos han sido proporcionados por la empresa Idealista* y contienen información sobre 189.923 viviendas situadas en las tres ciudades más grandes de España. Hasta la fecha, este producto de datos es uno de los mayores conjuntos abiertos de microdatos georreferenciados del mercado de la vivienda del mundo. El más similar en términos de desagregación geográfica y tamaño de la muestra es el conjunto de datos de Song et al. (2021), que incluye transacciones para cuatro ciudades de Corea del Sur, a saber, Busan ($n = 61,152$), Daegu ($n = 32,363$), Daejeon ($n = 21,114$) y Gwangju ($n = 25,984$). *{idealista18}* es sin duda el mayor producto de datos abiertos de este tipo en España.

El conjunto de datos ha sido suministrado directamente por Idealista, por lo que está limpio y libre de errores de descarga. No obstante, para cumplir la legislación sobre datos, hemos enmascarado los precios aplicando una pequeña cantidad de ruido aleatorio que no sesgará los principales resultados derivados de su uso.

Este conjunto de microdatos puede utilizarse para comparar nuevos métodos de forma reproducible (por ejemplo, Rey-Blanco et al., 2024). Los investigadores aplicados y teóricos sobre métodos de tasación y valoración masiva de bienes inmuebles podrían utilizar este conjunto de datos para comparar canónicamente el rendimiento de sus modelos comparables y hedónicos propuestos, entre otros. Los datos también pueden utilizarse para estudiar la segmentación de los submercados de la

vivienda y temas relacionados, como el impacto de las zonas suburbanas en los precios de la vivienda. Los listados se han enriquecido con

∗Idealista es el principal portal inmobiliario de España, y está presente en otros países del sur de Europa como Italia y Portugal.

información oficial del catastro español junto con otras características geográficas relevantes, como la distancia a puntos de interés urbano. En cualquier caso, los datos podrían ampliarse fácilmente uniendo espacialmente otros conjuntos de datos que contengan información a nivel administrativo, de sección censal o de código postal.

El conjunto de datos se distribuye como un paquete de R y se puede acceder a él desde un repositorio público de Github[†]. El producto de datos abiertos está disponible bajo la licencia Open Database License. En aras de la transparencia, también compartimos el proceso de enmascaramiento aplicado a los datos originales en el mencionado repositorio de Github.

Descripción de los datos

El producto de datos abiertos *{idealista18}* es un paquete R compuesto por diez objetos, tres objetos para cada una de las tres principales ciudades españolas: Barcelona, Madrid y Valencia. Para cada ciudad se han incluido en el paquete listados de viviendas, polígonos de barrios y un conjunto de puntos de interés. Hay además un objeto de datos con el número de viviendas por distrito (una colección de barrios), según el catastro español. El proveedor de datos (idealista.com) es un portal inmobiliario líder en España, a la par que su competidor más cercano, Fotocasa. Otros portales de anuncios más pequeños son Habitaclia y Milanuncios, este último centrado casi exclusivamente en anunciantes particulares (es decir, no profesionales). En septiembre de 2021, según datos de Similarweb (sitio especializado en la comparación del volumen de tráfico de los sitios), hubo un total de 103 millones de páginas vistas en portales inmobiliarios en España. De ellos, el 94% del tráfico se concentró en cuatro portales, de los que idealista fue el más importante, seguido de Fotocasa, Habitaclia y Pisos.com. El tráfico también está muy concentrado, siendo idealista.com el líder con diferencia con 58,6 millones de visitas mensuales (57% del tráfico total) frente a su competidor más cercano con 19,9 millones de visitas (19,3 % del tráfico total).

Gracias a su cuota de anuncios, idealista.com cubre bastante bien todos los segmentos del mercado español, incluidos los anunciantes particulares y profesionales. Este conjunto de datos incluye información sobre los precios de venta, por lo que representa la situación del mercado desde la perspectiva de los precios de venta. Se trata de un compromiso necesario en el caso que nos ocupa, ya que los precios de venta reales no están a disposición del público y sólo se **puede** acceder a la información pagando elevadas tasas al Colegio de Registradores. Sin embargo los precios de lista reflejan bastante bien la realidad (de las transacciones) de los mercados inmobiliarios, y pueden establecerse correlaciones entre los precios de lista idealistas y los precios de transacción véase Banco de Espan˜a. Las transacciones oficiales y los precios de venta pueden considerarse complementarios y son de gran interés a la hora de estudiar las diferencias entre los precios de venta y transacción o la relación entre las variables de demanda de los sitios de anuncios (es decir, contactos o visitas a los anuncios) y las diferencias de precios.

Para proporcionar algo de contexto sobre la cobertura del conjunto de datos *{idealista18},* la Tabla 1 muestra el número de listados con respecto al parque residencial total en cada ciudad en 2018. Como se observa en la tabla, el número de

listados oscila entre el 6,1% del total de inmuebles (en Madrid) y el 8,1% (en Valencia). Información procedente del Instituto Nacional de

[†]https://github.com/paezha/idealista18

Estadística‡ muestra que el número de anuncios del paquete {idealista18} corresponde al 81,3% de las transacciones inmobiliarias registradas en Barcelona, al 80,8% en Madrid y al 91,1% en Valencia.

Tabla 1. Total inmuebles y transacciones tres ciudades españolas. Año 2018

Ciudad	Total de propiedades (TP)	Total de transacciones (TT)	Listados (L)	L/TP	TT/L
Barcelona	789,740	56,012	61,329	7.8%	81.3%
Madrid	1,545,397	76,603	94,802	6.1%	80.8%
Valencia	416,004	30,615	33,593	8.1%	91.1%
Total	2,751,141	163,230	189,724	6.9%	86.0%

Fuentes

Total propiedades (P): Ministerio Espan˜ol de Hacienda y Funcio'n Pu'blica

T o t a l transacciones (T): Instituto Nacional de Estad'ıstica

Las siguientes subsecciones describen los objetos de datos. También se puede consultar una descripción completa de los datos en la sección de ayuda del paquete. Hemos intentado, en la medida de nuestras posibilidades, cumplir con los principios FAIR relativos a los datos de investigación (Wilkinson et al., 2016): tras la publicación, el conjunto de datos tiene un identificador de objeto digital persistente; la publicación como un artículo de datos hace que los datos sean localizables; los datos y metadatos se empaquetan juntos y los protocolos para archivos de ayuda en el ecosistema `R` significan que la documentación se puede buscar fácilmente; la distribución como un paquete `R` significa que solo se necesita software abierto para acceder a los datos; y un repositorio público documenta todos los procesos de datos seguidos para generar el producto de datos abiertos distribuido.

Listados de viviendas

El listado de cada ciudad incluye un conjunto de características de cada vivienda publicadas en la web inmobiliaria idealista. El listado incluido en el paquete *{idealista18}* son objetos de características simples (sf) (Pebesma, 2018) con geometría puntual en latitud y longitud. El nombre de cada objeto sf con el listado de viviendas es el nombre de la ciudad seguido de ' Sale' (por ejemplo, `Madrid Sale`). Cada objeto de datos incluye un total de 42 variables y el conjunto completo de listados correspondientes a los cuatro trimestres de 2018 (Q1 a Q4). La Tabla 2 muestra el número de anuncios de venta de viviendas incluidos en el conjunto de datos para cada ciudad y trimestre. Los recuentos de registros para cada ciudad son: 94.815 listados para Madrid, 61.486 para Barcelona y 33.622 para Valencia. Nótese que la misma vivienda puede encontrarse en más de un periodo cuando una propiedad puesta en venta en un trimestre se vendió en un trimestre posterior. La variable ASSETID, incluida en los objetos sf, es el identificador único de la vivienda.

Cada registro del listado de viviendas contiene un conjunto de características interiores facilitadas por los anunciantes en la web de Idealista (por ejemplo, precio, superficie, número de dormitorios, características básicas, etc.), incluyendo una ubicación aproximada de la vivienda (la ubicación exacta se ha enmascarado, tal y como se describe en el apartado Enmascaramiento de los precios). En el Cuadro 3 se

enumeran los

[‡]https://www.ine.es/jaxiT3/Tabla.htm?t=6150&L=1

Ciudad	Primer	Segundo	Tercero	Cuarto	Total anuncios
Barcelona	17826	7951	12375	23334	61486
Madrid	21920	12652	15973	44270	94815
Valencia	9305	4655	5644	14018	33622

Tabla 2. Número de anuncios de viviendas por ciudad y trimestre.

principales variables interiores con una breve descripción y el valor medio de cada variable. Los listados de viviendas se enriquecieron con una serie de atributos adicionales procedentes del catastro español (Registro Central del Catastro, 2021). La información catastral se describe en la Tabla 3, con el prefijo CAD en el nombre de la variable. Las características catastrales se asignaron aplicando las características de la parcela más cercana a las coordenadas. El año de construcción de la vivienda (CONSTRUCTIONYEAR) facilitado por el anunciante se revisó ya que el año de construcción es introducido en la web por los usuarios, por lo que está sujeto a errores y datos incompletos (40% de datos perdidos). Para resolver este problema, se incluyó una variable alternativa (CADCONSTRUCTIONYEAR), que asigna el año de construcción catastral de la parcela catastral más cercana siempre que el valor estuviera pendiente (la fecha fuera posterior a la fecha de publicación o el año de construcción fuera anterior a 1500) o cuando faltara el valor facilitado por el anunciante.

Adicionalmente, se incluyó en el objeto sf la distancia de cada vivienda a varios puntos de interés: distancia al centro de la ciudad, distancia a la estación de metro más cercana y distancia a una calle principal (La Diagonal para Barcelona, La Castellana para Madrid y Blasco Ibáñez para Valencia). Las últimas filas de la Tabla 3 muestran los valores medios de estas variables.

Descripción de	VariableSort	Barcelona	Madrid	Valencia
PRECIOPrecio de venta	395770.58	396110.11	199678.31	
Precio	unitario por m^2 (euros)	4044.86	3661.05	1714.54
SUPERFICIE	DEAREAS CONSTRUIDAS (m^2)	95.46	101.40	108.95
	NÚMERO DE HABITACIONESNúmero de habitaciones			2.86
2.58		3.07		
	NÚMERO DE BAÑOSNúmero de baños	1.52	1.59	1.59
AÑO DE CONSTRUCCIÓNAño de construcción (anunciante)		1952.58	1964.69	1969.43
CADCONSTRUCTIONYEARAño de construcción (catastro)		1952.19	1965.70	1970.55
	CADMAXBUILDINGFLOORMax suelo de construcción			6.85
6.38		7.04		
CADDWELLINGCOUNTCuenta de viviendas en el edificio		28.56	39.19	36.83
CALIDAD CADASTRALIDCalidad catastral. 0 Mejor-10 Peor		4.31	4.85	5.34
DISTANCIA AL CENTRO DE LA CIUDAD Distancia al centro de la ciudad			2.80	4.49
		2.09		
DISTANCIA AL	METRODistancia a la estación de metro	0.27	0.48	0.64
DISTANCIA A (CALLE PRINCIPAL) Distancia a calle principal		1.77	2.68	2.07

Tabla 3. Lista, descripción ordenada y media de las principales variables cuantitativas incluidas en el listado de viviendas para las tres ciudades españolas. Véase la sección de

ayuda del paquete *{IDEALISTA18}* para más detalles y definiciones formales. Algunas variables se han excluido de esta tabla para ahorrar espacio. Consulte la lista completa en el paquete.

Además de las variables enumeradas en el cuadro 3, el objeto sf incluye un conjunto de variables ficticias con información sobre las características básicas de la vivienda. El cuadro 4 muestra las variables más relevantes incluidas en el objeto sf.

Descripción de	VariableSort	Barcelona	Madrid	Valencia	
	HASTERRACE=1 si tiene terraza	0.33	0.36	0.25	
	HASLIFT=1 si tiene ascensor	0.74	0.70	0.79	
	HASAIRCONDITIONING=1 si tiene aire acondicionado			0.47	0.45
	0.47				
	HASPARKINGSPACE=1 si tiene aparcamiento		0.08	0.23	0.17
	HASNORTHORIENTATION=1 si tiene orientación norte			0.13	0.11
	0.13				
	HASSOUTHORIENTATION=1 si tiene orientación sur			0.31	0.24
	0.19				
	HASEASTORIENTATION=1 si tiene orientación este			0.24	0.20
	0.25				
	HASWESTORIENTATION=1 si tiene orientación oeste			0.16	0.15
	0.15				
	HASBOXROOM=1 si tiene trastero	0.12	0.26	0.13	
	HASWARDROBE=1 si tiene armario	0.30	0.57	0.53	
	HASSWIMMINGPOOL=1 si tiene piscina		0.03	0.15	0.07
	HASDOORMAN=1 si tiene portero	0.08	0.25	0.05	
	HASGARDEN=1 si tiene jardín	0.04	0.18	0.06	
	ISDUPLEX=1 si es dúplex	0.03	0.03	0.02	
	ISSTUDIO=1 si es estudio	0.02	0.03	0.01	
	ISINTOPFLOOR=1 está en el último piso		0.02	0.02	0.01
BUILTTYPEID_	1=1 si es nueva construcción	0.01	0.03	0.03	
BUILTTYPEID_	2=1 es de segunda mano para restaurar	0.17	0.19	0.13	
BUILTTYPEID_3	es de segunda mano en buen estado	0.82	0.78	0.83	

Tabla 4. Lista de variables ficticias, descripción de la clasificación y ratios de viviendas con las características específicas. Consulte la sección de ayuda del paquete **idealista18** R para obtener más detalles y definiciones formales. Algunas variables ficticias se han excluido de esta tabla para ahorrar espacio

Polígonos de barrio

El segundo objeto de datos incluido en *{idealista18}* son los barrios de las tres ciudades como polígonos. Hay un objeto sf para cada ciudad con el nombre de la ciudad y el sufijo ' Polígonos'. La columna izquierda de la Figura 1 muestra los mapas de cuantiles del número de viviendas en el listado para los distintos barrios de las tres ciudades. Los barrios se basan en los límites oficiales pero ligeramente modificados por Idealista[§]. En términos prácticos, podemos suponer que son los mismos ya que el sitio web combina zonas cuando hay pocos anuncios para esa zona. En el caso de Madrid, combinaron cuatro zonas en dos. Hay un total de 69 barrios en Barcelona, 135 en Madrid y 73 en Valencia.

El objeto sf incluye un identificador único (LOCATIONID) y el nombre del barrio (LOCATIONNAME).

El número total de viviendas está disponible en el catastro español agregado por distritos (los distritos son grupos de barrios). La columna de la derecha del gráfico 1

muestra el porcentaje de viviendas censadas en relación con el número total de viviendas por distrito en cada uno de los distritos.

§ Para realizar esta división se utilizan dos criterios. Si un área es lo suficientemente pequeña y similar a otra, las dos áreas se fusionan y, si el área oficial no es homogénea, se divide en una serie de nuevos polígonos.

de las tres ciudades. Esto da una idea de lo activos que estuvieron los mercados inmobiliarios residenciales en diferentes partes de cada ciudad en 2018.

Puntos de interés

El último objeto de datos incluido en el paquete es un conjunto de Puntos de Interés de cada ciudad como objeto de la clase lista. El nombre de la lista incluye el nombre de la ciudad con el sufijo ' POIS'. Estas listas incluyen tres elementos (i) las coordenadas del centro de la ciudad, el distrito central de negocios; (ii) un conjunto de coordenadas que definen la calle principal de cada ciudad ; y (iii) las coordenadas de las estaciones de metro.

Enmascarar los precios

Para cumplir la normativa española, se modificaron ligeramente tres variables para garantizar el anonimato. Se aplicó un proceso de enmascaramiento a los precios de venta y a la ubicación (coordenadas).

En cuanto a los precios de venta, los valores originales se ofuscaron con la adición o sustracción de un porcentaje aleatorio de sus valores originales, que oscilaba entre -2,5 y +2,5. Dado que los precios de venta suelen ser múltiplos de 1.000, tras la primera modificación de precios, los precios se ajustaron a múltiplos de 1.000.

Para entender las implicaciones de este proceso de enmascaramiento, utilizamos algunos resultados estándar del álgebra de variables aleatorias. Los precios enmascarados P vienen dados por:

$$P = RP - \epsilon$$

donde RP son los precios originales (brutos), y ϵ es una variable aleatoria extraída de la distribución uniforme con parámetros $a = 0{,}975$ y $b = 1{,}025$:

$$f(\epsilon) = \begin{cases} \frac{1}{b-a} & \text{para } a \le \epsilon \le b \\ 0 & \text{de lo} \end{cases}$$

contrario La expectativa de ϵ dados estos parámetros

es:

$$E[\epsilon] = \frac{a+b}{2} = \frac{0.975 + 1.025}{2} = 1$$

y la varianza de ϵ es:

$$V[\epsilon] = \frac{(b-a)^2}{12} = \frac{1}{4800}$$

Por lo tanto, la expectativa de los precios enmascarados es:

$$E[P] = E[RP - \epsilon] = E[RP] - E[\epsilon] = E[RP]$$

En otras palabras, los precios enmascarados P son una versión no sesgada de los precios brutos RP .

Considerando que RP y ϵ son independientes, la varianza de los precios enmascarados es la siguiente:

$$V[P] = V[RP - \epsilon] = V[RP] - V[\epsilon] + V[RP] - (E[\epsilon])^2 + V[\epsilon] - (E[RP])^2$$

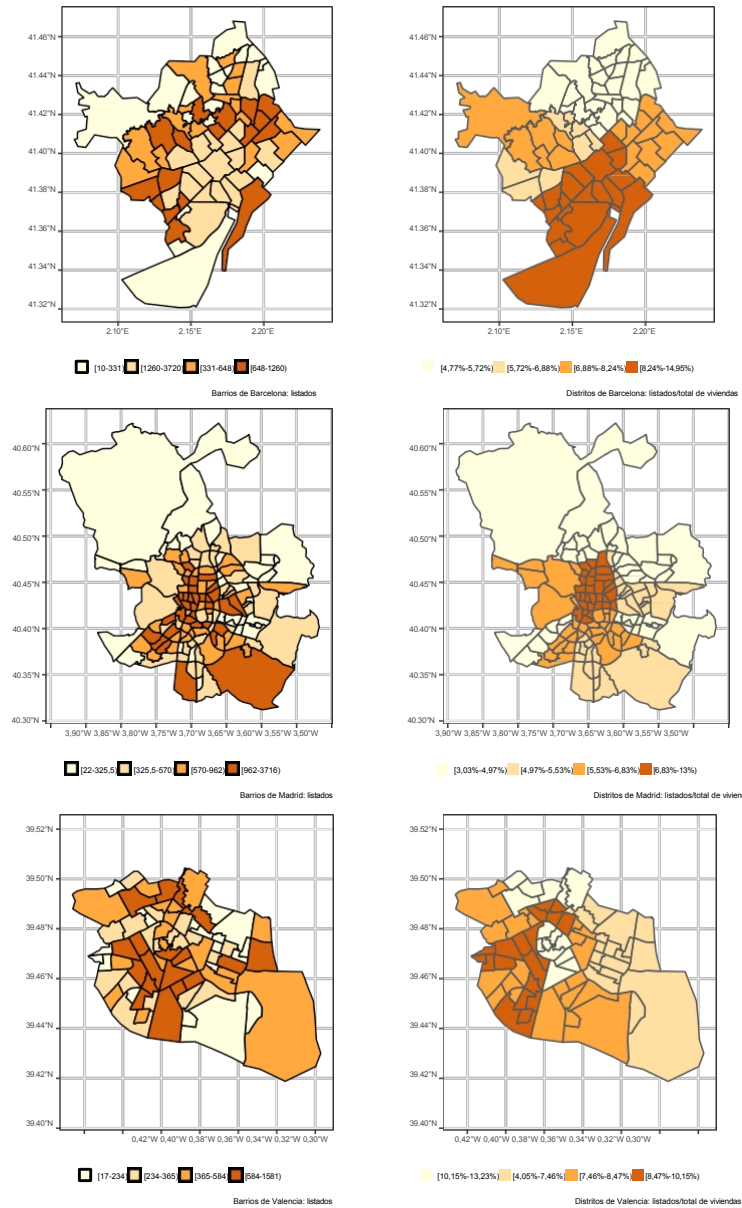

Figura 1. Listados por barrios (columna izquierda) y porcentaje de listados respecto al total de viviendas (columna derecha). Barcelona (arriba), Madrid (centro) y Valencia (abajo).

Como E[ϵ] = 1, tenemos que:

$$V[P] = V[RP] - V[\epsilon] + V[RP] + V[\epsilon] - (E[RP])^2 = V[RP] - (1 + V[\epsilon]) + V[\epsilon] - (E[RP])^2$$

Resolviendo para V[RP], y sustituyendo la expectativa de los precios brutos por su versión insesgada (E[P]), se obtiene la varianza de los precios brutos:

$$V[RP] = \frac{V[P] - \frac{1}{4800} \cdot (E[P])^2}{1 + \frac{1}{4800}}$$

El cuadro 5 presenta la media y la desviación típica (es decir, la raíz cuadrada de la varianza) de los precios del paquete, así como la desviación típica (de nuevo, la raíz cuadrada de la varianza) de los precios brutos calculados mediante la fórmula anterior. Esto se hace para cada trimestre y para todo el año. La última columna del cuadro puede interpretarse como un factor de inflación. Vemos que la varianza de los precios enmascarados está inflada con respecto a la varianza de los valores originales en menos de un 1% en todos los casos examinados. Los usuarios pueden utilizar la fórmula anterior para calcular la inflación de la varianza si utilizan submuestras distintas de las mostradas, para evaluar los posibles efectos del enmascaramiento (por ejemplo, al calcular intervalos de confianza).

Con respecto a la ubicación de la vivienda, se aplicó un proceso de enmascaramiento espacial para mantener las propiedades espaciales del conjunto de datos original. Las coordenadas de cada listado se desplazaron mediante un procedimiento estocástico. Los listados se registraron utilizando coordenadas contenidas en círculos de desplazamiento máximo y mínimo, como se muestra en la figura 2 (izquierda). Para preservar la inclusión en un vecindario, el procedimiento de enmascaramiento espacial se restringió para garantizar que las coordenadas enmascaradas permanecieran en el vecindario original del listado.

Datos: todos los anuncios de idealista
Resultado: todos los listados idealistas con coordenadas enmascaradas
1 inicialización;
2 **para** *cada listado L* **do**
3 tomar la localización geográfica de L como (X, Y) **repetir**
4 tomar un ángulo aleatorio α de 0 a 360 grados tomar una distancia R como valor aleatorio de 30 a 60 metros determinar un nuevo punto (X', Y') calculado como un punto situado R con el ángulo α
5 **hasta** *esta condición de parada*;
6 establezca (X', Y') como nueva ubicación para el listado L
7 **final**

Algoritmo 1: Proceso de desplazamiento de coordenadas con fines de enmascaramiento

El algoritmo 1 desplaza iterativamente las coordenadas de cada listado con una distancia mínima y una distancia máxima con la restricción de que las nuevas

coordenadas no caigan en un vecindario diferente. De este modo se garantiza la conservación de los atributos de vecindad.

Cuadro 5. Inflación de la varianza de los precios enmascarados con respecto a los precios brutos

Ciudad	Periodo	media(P)	sd(P)	sd(RP)	sd(P)/sd(RP)
BARCELONA	Q1	405,166.8	308,623.8	308,536.2	1.00057
	Q2	388,053.5	252,520.5	252,432.1	1.0007
	Q3	382,692.5	268,880.9	268,796.1	1.00063
	Q4	398,157.8	275,459.3	275,370.6	1.00064
	2018	395,770.6	281,554.8	281,467.5	1.00062
MADRID	Q1	404,960.8	447,935.3	447,850.5	1.00038
	Q2	367,527.5	383,093.9	383,017.2	1.0004
	Q3	365,467.2	359,118.3	359,042.2	1.00042
	Q4	410,952.7	428,852.7	428,767	1.0004
	2018	396,110.1	417,074.4	416,991.8	1.0004
VALENCIA	Q1	204,836.5	187,957.4	187,914.6	1.00046
	Q2	176,661	141,002.4	140,964.6	1.00054
	Q3	188,189.4	167,146	167,106.6	1.00047
	Q4	208,523.5	183,452.7	183,409	1.00048
	2018	199,678.3	177,156	177,114.1	1.00047

Nota:
P: precios enmascarados;
RP: precios en bruto;
sd: desviación típica (raíz cuadrada de la varianza)

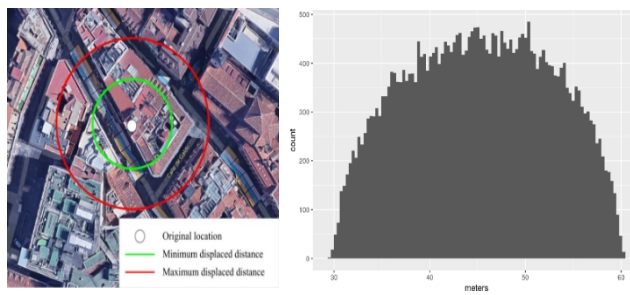

Figura 2. (Izquierda) Coordenadas de enmascaramiento. Rango espacial. (Derecha) Desplazamiento de coordenadas en metros (Valencia).

La figura 2 (derecha) muestra el histograma de los desplazamientos en metros para todos los listados de la ciudad de Valencia. La distancia media entre las coordenadas originales y las enmascaradas es de 45 metros.

Conclusión

Este artículo describe un producto de datos de un conjunto de microdatos georreferenciados de las tres ciudades más grandes de España. Este producto de datos puede ser valioso para apoyar la investigación sobre los mecanismos del mercado y los precios de la vivienda. Los investigadores pueden aplicar modelos hedónicos con efectos espaciales, identificar submercados de vivienda o experimentar con técnicas de aprendizaje automático. El producto de datos también puede utilizarse para propuestas educativas y actividades docentes. Por lo que sabemos, se trata del mayor conjunto de datos de este tipo disponible públicamente que, además, está listo para el análisis y completamente documentado.

Declaración de intereses en conflicto

Los autores Uno y Dos trabajan para Idealista. Se les ha concedido permiso para compartir los datos presentados en este artículo. Ninguno de los autores tiene intereses económicos o relaciones personales que hayan influido, o pudiera percibirse que han influido, en el trabajo presentado en este artículo.

Agradecimientos

Los autores desean dar las gracias a Alessandro Galesi por su apoyo en la revisión del trabajo y a Juan Ramo'n Selva por la recogida y limpieza de los datos espaciales. Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad de España con la subvención PID2019-107800GB-100, pero no ha sido financiado por ninguno de los consejos de investigación de Canadá.

Referencias

Arribas-Bel D, Alvanides S, Batty M, Crooks A, See L y Wolf L (2021a) Datos/código urbano: Una nueva sección de EP-b. *Environment and Planning B: Urban Analytics and City Science* 48(9): 2517-2519. DOI:https://doi.org/10.1177/23998083211059670.

Arribas-Bel D, Green M, Rowe F y Singleton A (2021b) Open data products-a framework for creating valuable analysis ready data. *Revista de Sistemas Geográficos* 23(4): 497-514. DOI: https://doi.org/10.1007/s10109-021-00363-5.

Bonifaci P y Copiello S (2015) Mercado inmobiliario y eficiencia energética de los edificios: Datos para un enfoque de valoración masiva. *Data in Brief* 5: 1060-1065. DOI:https://doi.org/https: //doi.org/10.1016/j.dib.2015.11.027.

Brunsdon C y Comber A (2021) Opening practice: supporting reproducibility and critical spatial data science. *Revista de Sistemas Geográficos* 23(4): 477-496. DOI:https://doi.org/10.1007/ s10109-020-00334-2.

Del Giudice V, De Paola P y Forte F (2018) Precios de alquiler de la vivienda: Datos de un área urbana central de nápoles (italia). *Datos en breve* 18: 983-987.

DOI:https://doi.org/10.1016/j.dib.2018.03.121.

Fuerst F y Haddad MFC (2020) Real estate data to analyse the relationship between property prices, sustainability levels and socio-economic indicators. *Data in Brief* 33: 106359. DOI: https://doi.org/10.1016/j.dib.2020.106359.

Li H, Wei YD, Wu Y y Tian G (2019) Análisis de los precios de la vivienda en shanghai con datos abiertos: Amenidad, accesibilidad y estructura urbana. *Ciudades* 91: 165-179. DOI:https://doi.org/10.1016/ j.cities.2018.11.016.

Lo'pez FA, Chasco C y Gallo JL (2015) Exploring scan methods to test spatial structure with an application to housing prices in m adrid. *Papers in Regional Science* 94(2): 317-346. DOI: https://doi.org/10.1111/pirs.12063.

Pa'ez A (2021) Ciencias espaciales abiertas: una introducción. *Revista de Sistemas Geográficos* 23(4): 467-476. DOI:https://doi.org/10.1007/s10109-021-00364-4.

Pebesma E (2018) Características simples para R: Soporte estandarizado para datos vectoriales espaciales. *The R Journal* 10(1): 439-446. DOI:https://doi.org/10.32614/RJ-2018-009.

Registro Central del Catastro (2021) https://www.sedecatastro.gob.es/ .

Rey-Blanco D, Arbue's P, Lo'pez FA y Pa'ez A (2024) Using machine learning to identify spatial market segments. a reproducible study of major spanish markets. *Environment and Planning B: Urban Analytics and City Science* 51(1): 89-108. DOI:https://doi.org/10.1177/ 23998083231166952.

Solano Sa'nchez MA' , Nu'n˜ez Tabales JM, Caridad y Ocerin JM, Santos JAC y Santos MC (2019). Conjunto de datos de precios diarios de alquileres vacacionales en un destino de turismo cultural. *Datos en breve* 27: 104697. DOI:https://doi.org/10.1016/j.dib.2019.104697.

Song Y, Ahn K, An S y Jang H (2021) Hedonic dataset of the metropolitan housing market - cases in south korea. *Data in Brief* 35: 106877. DOI:https://doi.org/10.1016/j.dib.2021. 106877.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, González-Beltrán A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 'T Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J y Mons B (2016) The fair guiding principles for scientific data management and stewardship. *Datos científicos* 3(1): 160018. DOI:https://doi.org/10.1038/sdata.2016.18.