A geo-referenced micro-data set of real estate listings for Spain's three largest cities

Journal Title
XX(X):1–13

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Abstract

This article ~~shares~~ presents an open data product with ~~big~~ large geo-referenced micro-data sets of 2018 real estate listings in Spain. These data were originally published on the idealista.com real estate website. The observations were obtained for the three largest cities in Spain: Madrid (n = 94,815 observations), Barcelona (n = 61,486 observations), and Valencia (n = 33,622 observations). The data sets include the coordinates of properties (latitude and longitude), asking prices of each listed dwelling, and several variables of indoor characteristics. The listings were enriched with official information from the Spanish cadastre (e.g., building material quality) plus other relevant geographical features, such as distance to urban points of interest. Along with the real estate listings, the data product also includes neighborhood boundaries for each city. The data product is offered as a fully documented ~~R~~ R package and is available for scientific and educational purposes, particularly for geo-spatial studies

Keywords

Housing prices; hedonic price analysis; idealista.com; geo-referenced data; point-level data; open data; Spain

Introduction

Interest in the characteristics of the housing market and housing prices has been a growing area of research in recent decades, generating a vast amount of theoretical and empirical literature. Including the spatial component to analyze the real estate market and incorporating geographic variables has significantly improved the understanding

of this market. But to really understand the characteristics of the housing market: to this end, it is essential to have information/data at the point level. Therefore, it is becoming common for spatial analysis of urban environments to be developed with geo- referenced use micro-datasets (López et al., 2015). However, the availability of this type of open data at the point level is limited, and not many data sets contain latitude/longitude coordinates for each dwelling. , geo-referenced as points (López et al., 2015). In some cases, researchers have had to resort to web scraping webscraping processes to obtain the large volumes of information that permit robust analyses (**??**Li et al., 2019; López et al., 2015). These web scraping processes can include missing data , data for research (e.g., Li et al., 2019; López et al., 2015). Alas, webscraping is a chancy process prone to download errors, duplicate missing data, duplicated records, etc. Furthermore, the authors of this research do not generally share the data sets.

We are also witnessing researchers do not always share their webscrapped datasets, which limits reproducibility of their research. As we witness a growing interest in open data in geography and openness and reproducibility in geographical data science (Arribas-Bel et al., 2021a; Páez, 2021; Brunsdon and Comber, 2021), it becomes increasingly urgent to have open data science (Arribas-Bel et al., 2021a,b) using reproducible or replicable research (Páez, 2021; Brunsdon and Comber, 2021). But to work openly in science, it is necessary to have free software and open data . While great efforts have been made to make free software availableto researchers (e.g., R or Python), data regarding products to support research (Arribas-Bel et al., 2021b).

Some researchers have already responded to this need for publicly available, geo-referenced datasets to support open, reproducible research. A few such datasets are now available to support the analysis of real estate marketstend to be treated as confidential, and there are still few open micro-data sets of housing markets available (**?**). , but they are sometimes geo-referenced at the level of large geographical zones, such as Fuerst and Haddad (2020), an open dataset that includes $n = 4,201$ property prices geocoded to the level of nine regions in England and Wales. Other datasets are geo-coded as points, including Bonifaci and Copiello (2015), which includes $n = 1,042$ observations for Padua, in Italy; Del Giudice et al. (2018), who share a dataset with $n = 576$ observations relating to rental prices in Naples, Italy; and Solano Sánchez et al. (2019) present a dataset with $n = 1,623$ daily rental prices in Seville, Spain.

To contribute to the growing inventory of international micro-data sets of real estate markets, this paper introduces an open micro-data set of data product with geo-referenced dwelling listings, called {idealista18}. The data have been provided by the Idealista company* and contain information about 189,923 dwellings located in Spain's three largest cities. To date, this data product is the biggest one of the largest open geo-referenced micro-data set of the housing market in Spain. Moreover, the data the

*Idealista is the major real estate listing website in Spain, and present in other southern european countries as Italy and Portugal

world. The most similar in terms of geographical disaggregation and sample size is the dataset of Song et al. (2021) which includes transactions for four cities in South Korea, namely Busan ($n = 61,152$), Daegu ($n = 32,363$), Daejeon ($n = 21,114$), and Gwangju ($n = 25,984$). {idealista18} is certainly the largest open data product of its kind in Spain.

The data set has been supplied directly by Idealista, and therefore is clean and free ~~of~~ from download errors. However, ~~in order to be able to share publicly the data we have slightly modified some variables by applying some random noise which~~ to comply with data legislation, we have masked the prices by applying a small amount of random noise that will not bias the main results derived from its usage.

This micro-data set ~~is expected to be used as a benchmark dataset to test empirical~~ can be used to benchmark new methods in a reproducible fashion (e.g., Rey-Blanco et al., 2023). Applied and theoretical researchers on real estate mass appraisal and valuation methods might use this dataset to canonically compare the performance of their proposed comparable and hedonic models, among others. The data can also be used to study the segmentation of housing submarkets and related topics such as the impact of suburban areas on house prices. The listings have been enriched with official information from the Spanish cadastre along with other relevant geographical features, such as distance to urban points of interest. In any case, the data might be easily extended by spatially joining other datasets that contain information at administrative, census tract, or postal code levels.

The data set is distributed as an ~~R package , name {idealista18}, which~~ R package and it can be accessed from a public Github repository[†]. The ~~database idealista18~~ open data product is made available under the Open Database License. For transparency, we also share the ~~randomisation~~ masking process applied to the original data in the aforementioned Github repository.

Data description

The open data product {idealista18} is an ~~R~~ R package composed of ~~nine~~ ten objects, three objects for each of the three main Spanish cities: Barcelona, Madrid, and Valencia. For each city, dwelling listings, neighborhood polygons, and a set of points of interest have been included in the ~~R package.~~ package. There is in addition a data object with the number of dwellings by district (a collection of neighborhoods), according to the Spanish cadastre.

The data provider (idealista.com) is a leading real estate portal in Spain, on par with its nearest competitor Fotocasa. Smaller listing portals include Habitaclia and Milanuncios, the latter focusing almost exclusively on individual (i.e., non-professional) advertisers. In September 2021, according to data from Similarweb (a site specialized in sites' traffic volume comparison), there were a total of 103 million page views on real estate portals in Spain. Of this, 94% of traffic was concentrated on four portals, of which idealista was the

[†]~~The direct URL to the GitHub repository is~~ `https://github.com/paezha/idealista18`

most important, followed by Fotocasa, Habitaclia, and Pisos.com. Traffic is also highly concentrated, and idealista.com was the the leader by far with 58.6 monthly million visits (57% of the total traffic) compared to its closest competitor with 19.9 million visits (19.3 % of total traffic).

As a result of its share of advertisments, idealista.com covers fairly well all segments of the Spanish market, including both individual and professional advertisers. This dataset includes information about listing prices, and therefore represents the market situation from the perspective of asking prices. This is a necessary compromise in the present case, since actual sale prices are not publicly available, and the information can only be accessed by paying high fees to Colegio de Registradores. However listing prices reflect quite well the (transaction) reality of real estate markets, and correlations between idealista listings and transaction prices can be established see Banco de España. Official transactions and asking prices can be taken as complementary and are of great interest when studying asking-transaction price gaps or the relation between listing site demand variables (i.e., ad contacts or ad views) and price gaps.

To provide some context about the coverage of the {idealista18} dataset Table 1 shows the number of listings with respect to the total residential stock in each city in 2018. As seen in the table, the number of listings ranges between 6.1% of the total number of properties (in Madrid) and 8.1% (in Valencia). Information from Instituto Nacional de Estadistica[‡] shows that the number of listings in the {idealista18} package correspond to 81.3% of recorded real estate transactions in Barcelona, 80.8% in Madrid, and 91.1% in Valencia.

Table 1. Total properties and transactionst three Spanish cities. Year 2018

City	Total properties (TP)	Total transactions (TT)	Listings (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%

Sources

Total properties (P): Ministerio Español de Hacienda y Función Pública

Total transactions (T): Instituto Nacional de Estadística

The following subsections describe ~~each object~~the data objects. A full description of the data is also available in the help section of the package. We have ~~strived~~ tried to the best of our ability to comply with FAIR principles regarding research data (Wilkinson et al., 2016): upon publication, the dataset has a persistent digital object identifier; publication as a data article makes the data findable; the data and metadata are packaged together and protocols for help files in the ~~R~~ R ecosystem mean that documentation is easily searchable; distribution as an ~~R~~ R package means that only open software is

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

needed to access the data; and a public repository documents all data processes followed to generate the distributed open data product.

Dwelling listings

The dwelling listing of listing for each city includes a set of characteristics for each dwelling published on the idealista real estate websiteas an ad. The dwelling listing has been . The listing included in the ‘{idealista18’ package as an sfobject (Pebesma, 2018) } package are simple features (sf) objects (Pebesma, 2018) with point geometry in latitude and longitude. The name of the sf object containing the dwelling listing includes each sf object with the list of dwellings is the name of the city , followed by ’_Sale’ (e.g., Madrid_Sale)and `Madrid_Sale`). Each data object includes a total of 42 variables . Each sf object includes and the complete set of listings corresponding to the four quarters of the year 2018. 2018 (Q1 through Q4). Table 2 shows the number of dwelling listing ads included in the data set for each city and quarter. The record counts for each city are: 94,815 listings for Madrid, 61,486 for Barcelona, and 33,622 for Valencia. Note that the same dwelling may be found in more than one period when a property listed for sale in one quarter was sold in a subsequent quarter. The variable ASSETID, included in the sf objects, is the unique identifier of the dwelling.

City\Quarter	First	Second	Thirdr Third	Fourth	Total ads
Barcelona	17826	7951	12375	23334	61486
Madrid	21920	12652	15973	44270	94815
Valencia	9305	4655	5644	14018	33622

Table 2. Number of dwelling listing ads for each city and quarter.

Each record of the dwelling listing contains a set of indoor characteristics supplied by the advertisers on the Idealista website (e.g., price, surface area, number of bedrooms, basic features, etc.), including the exact an approximated location of the dwelling (see the exact location has been masked, as described in Section Masking the prices). Table 3 lists the main indoor variables with a short description and the mean value of each variable. The dwelling listings were enriched with a number of additional attributes from the Spanish cadastre (Registro Central del Catastro, 2021). The cadastral information is described in Table 3, with the prefix CAD in the variable name. The cadastral features were assigned by applying the features of the nearest parcel to the coordinates. The year the dwelling was built (CONSTRUCTIONYEAR) given by the advertiser was revised since the year of construction is entered on the website by users, and it is therefore subject to errors and incomplete data (40% missing data). To resolve this issue, an alternative variable (CADCONSTRUCTIONYEAR) was included, assigning the cadastral construction year from the nearest cadastral parcel whenever the value was outstanding (date was after publication date or year of construction was before 1500) or when the value supplied by the advertiser was missing.

Additionally, the distance of each dwelling to three urban several points of interest was included in the sf object: distance to the city center, distance to the closest metro station,

and distance to a major street (La Diagonal for Barcelona, La Castellana for Madrid, and Blasco Ibañez for Valencia). The last rows of Table 3 show the mean values of these variables.

Variable	Sort Description	Barcelona	Madrid	Valencia
PRICE	Asking price	395770.58	396110.11	199678.31
UNITPRICE	Asking price per m^2 (euros)	4044.86	3661.05	1714.54
CONSTRUCTEDAREA	Surface (m^2)	95.46	101.40	108.95
ROOMNUMBER	Number of bedrooms	2.86	2.58	3.07
BATHNUMBER	Number of bathrooms	1.52	1.59	1.59
CONSTRUCTIONYEAR	Construction year (advertiser)	1952.58	1964.69	1969.43
CADCONSTRUCTIONYEAR	Construction year (cadastre)	1952.19	1965.70	1970.55
CADMAXBUILDINGFLOOR	Max build floor	6.85	6.38	7.04
CADDWELLINGCOUNT	Dwelling count in the building	28.56	39.19	36.83
CADASTRALQUALITYID	Cadastral quality. 0 Best-10 Worst	4.31	4.85	5.34
DISTANCE_TO_CITY_CENTER	Distance to city center	2.80	4.49	2.09
DISTANCE_TO_METRO	Distance to subway station	0.27	0.48	0.64
DISTANCE_TO_(MAINSTREET)	Distance to major street	1.77	2.68	2.07

Table 3. List, sort description, and mean of the main quantitative variables included in the dwelling listing for the three Spanish cities. See the help section in the ~~**idealista18** R~~ {idealista18} package for details and formal definitions. Some variables have been excluded from this table to save space. Check the full list in the ~~**idealista18**~~ package.

In addition to the variables listed in Table 3, the sf object includes a set of dummy variables with information about the basic characteristics of the dwelling. Table 4 shows the more relevant variables included in the sf object.

Neighboorhood polygons

The second ~~block of data included in the~~ 'data object included in {idealista18' ~~R package is the spatial features of~~ } is the neighborhoods in the three cities ~~divided into neighborhoods~~as polygons. There is an sf object for each city with the name of the city and the suffix '_Polygons'. The left column of Figure 1 shows the quantile maps of the number of dwellings in the listing for the different neighborhoods in the three cities. The neighborhoods are based on the official boundaries but slightly changed by Idealista[§]. In practical terms, we can assume they are the same since the website combines areas when there are few ads for that area. In the case of Madrid, they combined four areas into two.

~~Quantile maps of the number of dwellings in each neighborhood. Boundary for Barcelona (Left), Madrid (Center), and Valencia (Right).~~

There are a total of 69 neighborhoods in Barcelona, 135 in Madrid, and 73 in Valencia. The sf object includes a unique identifier (LOCATIONID) and the neighborhood name (LOCATIONNAME).

The total number of dwellings is available from the Spanish cadastre aggregated by district (districts are groups of neighborhoods). The right column in Figure 1 shows the

[§]There are two criteria used to make this division. If an area is small enough and similar enough to another, the two areas are merged, and, if the official area is not homogeneous, it is divided into a series of new polygons.

Variable	Sort Description	Barcelona	Madrid	Valencia
HASTERRACE	=1 if has terrace	0.33	0.36	0.25
HASLIFT	=1 if has lift	0.74	0.70	0.79
HASAIRCONDITIONING	=1 if has air conditioning	0.47	0.45	0.47
HASPARKINGSPACE	=1 if has parking	0.08	0.23	0.17
HASNORTHORIENTATION	=1 if has north orientation	0.13	0.11	0.13
HASSOUTHORIENTATION	=1 if has south orientation	0.31	0.24	0.19
HASEASTORIENTATION	=1 if has east orientation	0.24	0.20	0.25
HASWESTORIENTATION	=1 if has west orientation	0.16	0.15	0.15
HASBOXROOM	=1 if has boxroom	0.12	0.26	0.13
HASWARDROBE	=1 if has wardrobe	0.30	0.57	0.53
HASSWIMMINGPOOL	=1 if has swimmingpool	0.03	0.15	0.07
HASDOORMAN	=1 if has doorman	0.08	0.25	0.05
HASGARDEN	=1 if has garden	0.04	0.18	0.06
ISDUPLEX	=1 if is duplex	0.03	0.03	0.02
ISSTUDIO	=1 if is studio	0.02	0.03	0.01
ISINTOPFLOOR	=1 is on the top floor	0.02	0.02	0.01
BUILTTYPEID_1	=1 if is new contruction	0.01	0.03	0.03
BUILTTYPEID_2	=1 is second hand to be restored	0.17	0.19	0.13
BUILTTYPEID_3	=1 is second hand in good condition	0.82	0.78	0.83

Table 4. List of dummy variables, sort description, and ratios of dwellings with the specific characteristics. See the help section in the **idealista18** R package for details and formal definitions. Some dummy variables have been excluded from this table to save space

percentage of listed dwellings relative to the total number of dwelling by district in each of the three cities. This gives a sense of how active residential real estate markets were in different parts of each city in 2018.

Points of Interest

The last ~~block of data~~ data object included in the ~~data~~ package is a set of Points of Interest in each city as an object of the class list. The name of the list includes the name of the city with the suffix '_POIS'. These lists include three elements: (i) the coordinates of the city center, the central business district; (ii) a set of coordinates that define the main street of each city; and (iii) the coordinates of the metro stations.

Masking the prices

To comply with Spanish regulations, ~~two~~ three variables were slightly modified to provide a measure of anonymity. A masking process was applied to asking prices and location (coordinates).

In terms of the asking prices, the original values were obfuscated with the addition or subtraction of a random percentage of their original values, ranging from ~~-2.5% to +2.5%~~ -2.5 to $+2.5$. Since asking prices are usually multiples of 1,000, after the first price modification, the prices were aligned to multiples of 1,000.

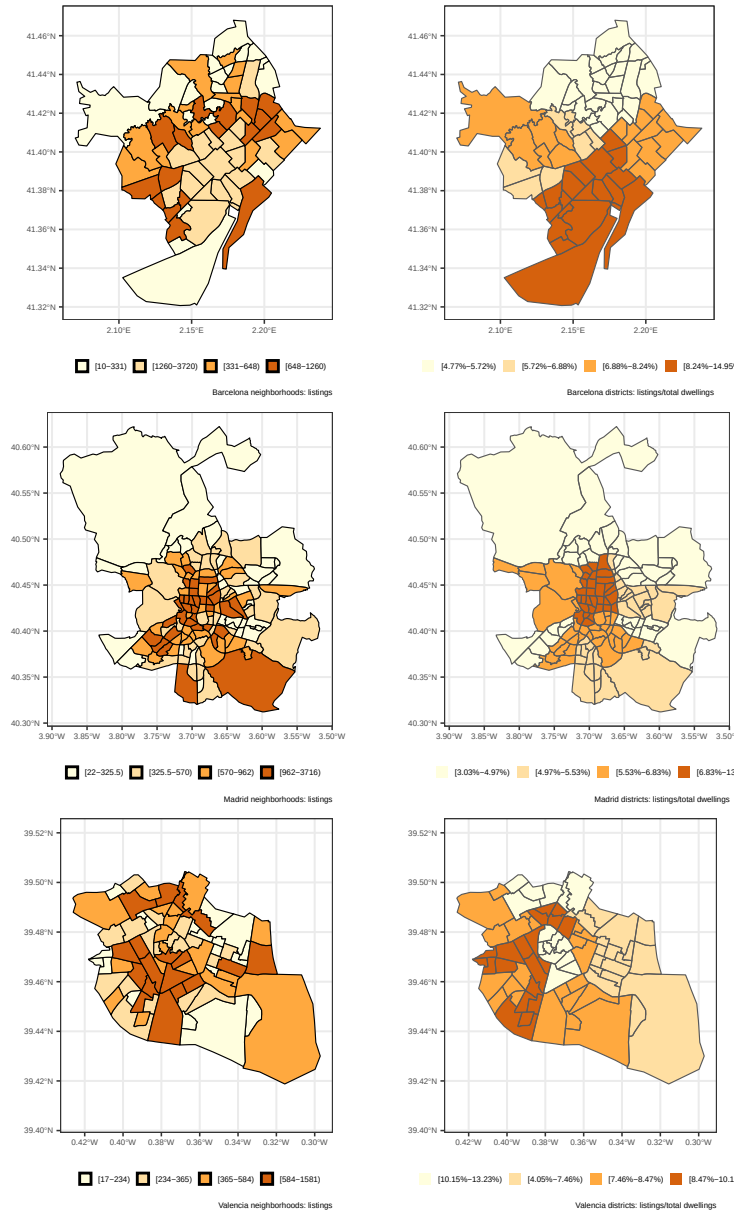

Figure 1. Listings by neighborhood (left column) and percentage of listings relative to total dwellings (right column). Barcelona (Top), Madrid (Center), and Valencia (Bottom).

~~initialization Coordinate displacement process for anonymisation purposes~~ To understand the implications of this masking process, we use some standard results from algebra of random variables. The masked prices P are given by:

$$P = RP \cdot \epsilon$$

where RP are the original (raw) prices, and ϵ is a random variable drawn from the uniform distribution with parameters $a = 0.975$ and $b = 1.025$:

$$f(\epsilon) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq \epsilon \leq b \\ 0 & \text{otherwise} \end{cases}$$

The expectation of ϵ given these parameters is:

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

and the variance of ϵ is:

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

Therefore, the expectation of the masked prices is:

$$\mathrm{E}[P] = \mathrm{E}[RP \cdot \epsilon] = \mathrm{E}[RP] \cdot \mathrm{E}[\epsilon] = \mathrm{E}[RP]$$

In other words, the masked prices P are an unbiased version of the raw prices RP.

Considering that RP and ϵ are independent, the variance of the masked prices is as follows:

$$\mathrm{V}[P] = \mathrm{V}[RP \cdot \epsilon] = \mathrm{V}[RP] \cdot \mathrm{V}[\epsilon] + \mathrm{V}[RP] \cdot (\mathrm{E}[\epsilon])^2 + \mathrm{V}[\epsilon] \cdot (\mathrm{E}[RP])^2$$

Since $\mathrm{E}[\epsilon] = 1$, we have that:

$$\mathrm{V}[P] = \mathrm{V}[RP] \cdot \mathrm{V}[\epsilon] + \mathrm{V}[RP] + \mathrm{V}[\epsilon] \cdot (\mathrm{E}[RP])^2 = \mathrm{V}[RP] \cdot (1 + \mathrm{V}[\epsilon]) + \mathrm{V}[\epsilon] \cdot (\mathrm{E}[RP])^2$$

Solving for $\mathrm{V}[RP]$, and replacing the expectation of the raw prices by its unbiased version ($\mathrm{E}[P]$), yields the variance of the raw prices:

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

Table 5 reports the mean and the standard deviation (i.e., the square root of the variance) of the prices in the package, and the standard deviation (again, the square

Table 5. Inflation of the variance of masked prices with respect to raw prices

City	Period	mean(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

Note:
P: masked prices;
RP: raw prices;
sd: standard deviation (square root of the variance)

root of the variance) of the raw prices calculated using the formula above. This is done for each quarter and for the full year. The last column of the table can be read as an inflation factor. We see that the variance of the masked prices is inflated with respect to the variance of the original values by less than 1% in all cases examined. Users can use the formula above to calculate the inflation of the variance if they use subsamples other than those shown, to assess the potential impacts of the masking (e.g., when computing intervals of confidence).

With respect to the location of the dwelling, a spatial masking process was implemented to maintain the spatial properties of the original data set. The coordinates of each listing were displaced using a stochastic procedure. The listings were recorded using coordinates contained in maximum and minimum displacement circles, as shown in Figure 2 (left). To preserve inclusion in a neighborhood, the spatial masking procedure was constrained to ensure that the masked coordinates remained in the original neighborhood of the listing.

Data: all idealista listings
Result: all idealista listings with masked coordinates
initialization; **for** *each listing* L **do**
  take geographical location of L as (X, Y) **repeat**
    take a random angle α from 0 to 360 degrees take a distance R as a random value from 30 to 60 meters determine a new point (X', Y') calculated as a point located R with the angle α
  until *this stop condition*;
  set (X', Y') as the new location for the listing L
end

Algorithm 1: Coordinate displacement process for masking purposes

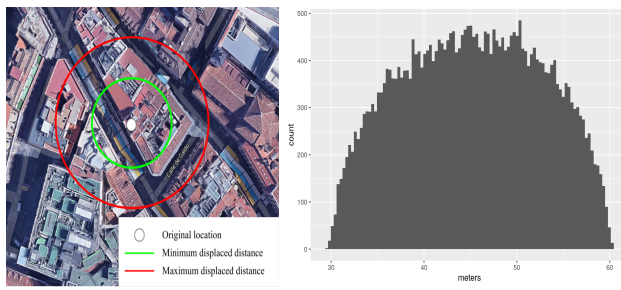

Figure 2. (Left) Masking coordinates. Spatial range. (Right) Coordinate displacement in meters (Valencia)

Algorithm 1 iteratively displaces the coordinates of each listing with a minimum distance and a maximum distance with the restriction that the new coordinates do not fall into a different neighborhood. This ensures that neighborhood attributes are preserved.

Figure 2 (right) shows the histogram of the displacements in meters for all the listings in the city of Valencia. The average distance between the original and masked coordinates is 45 meters.

Conclusion

This paper describes a data product of a geo-referenced micro-data set of Spain's three largest cities. This ~~is an excellent data product to help understand the complex mechanisms related to the~~ data product can be of value to support research into the mechanisms of housing market and housing prices. Researchers can apply hedonic models with spatial effects, ~~identifying housing submarketsor~~ identify housing submarkets, or experiment with machine learning techniques. The data product can also be used for educational proposes and teaching activities. To the best of our knowledge, this is the largest, publicly available data set of its type that is also analysis ready and fully documented.

Declaration of Competing Interest

~~Author One and author~~ Authors One and Two are employed by Idealista. They have been granted permission to share the data presented in this article. None of the authors have financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The authors wish to thank Alessandro Galesi for their support in the paper revision and Juan Ramón Selva for collecting and cleaning the spatial data. This work has been partially funded by the Spanish Ministry of Economy and Competitiveness Grants PID2019-107800GB-100, but it was not funded by any of Canada's research councils.

References

Arribas-Bel D, Alvanides S, Batty M, Crooks A, See L and Wolf L (2021a) Urban data/code: A new EP-b section. *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 and Singleton A (2021b) Open data products-a framework for creating valuable analysis ready data. *Journal of Geographical Systems* 23(4): 497–514. DOI: https://doi.org/10.1007/s10109-021-00363-5.

Bonifaci P and Copiello S (2015) Real estate market and building energy performance: Data for a mass appraisal approach. *Data in Brief* 5: 1060–1065. DOI:https://doi.org/https://doi.org/10.1016/j.dib.2015.11.027.

Brunsdon C and Comber A (2021) Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems* 23(4): 477–496. DOI:https://doi.org/10.1007/s10109-020-00334-2.

Del Giudice V, De Paola P and Forte F (2018) Housing rental prices: Data from a central urban area of naples (italy). *Data in Brief* 18: 983–987. DOI:https://doi.org/10.1016/j.dib.2018.03.121.

Fuerst F and 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 and Tian G (2019) Analyzing housing prices in shanghai with open data: Amenity, accessibility and urban structure. *Cities* 91: 165–179. DOI:https://doi.org/10.1016/j.cities.2018.11.016.

López FA, Chasco C and 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.

Páez A (2021) Open spatial sciences: an introduction. *Journal of Geographical Systems* 23(4): 467–476. DOI:https://doi.org/10.1007/s10109-021-00364-4.

Pebesma E (2018) Simple Features for R: Standardized Support for Spatial Vector Data. *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, Arbués P, López FA and Páez A (2023) Using machine learning to identify spatial market segments. a reproducible study of major spanish markets. *Environment and Planning B: Urban Analytics and City Science* 0(0): 23998083231166952. DOI:https://doi.org/10.1177/23998083231166952.

Solano Sánchez MÁ, Núñez Tabales JM, Caridad y Ocerin JM, Santos JAC and Santos MC (2019) Dataset for holiday rentals' daily rate pricing in a cultural tourism destination. *Data in Brief* 27: 104697. DOI:https://doi.org/10.1016/j.dib.2019.104697.

Song Y, Ahn K, An S and 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, Gonzalez-Beltran 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 and Mons B (2016) The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3(1): 160018. DOI:https://doi.org/10.1038/sdata.2016.18.