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

Journal Title
XX(X):1–7
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Abstract

This article shares an open data product with big 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 package and is available for scientific and educational purposes, particularly for geo-spatial studies

Keywords

Housing market; 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, 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 micro-data sets (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. In some cases, researchers have had to resort to web scraping processes to obtain the large volumes of information that permit robust analyses (Gupta et al., 2022; Arbia and Nardelli, 2020; Li et al., 2019; López et al., 2015). These web scraping processes can include missing data, download errors, duplicate records, etc. Furthermore, the authors of this research do not generally share the data sets.

We are also witnessing a growing interest in open data in geography and data science (Arribas-Bel et al., 2021a,b) using reproducible or replicable research (Páez, 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 available to researchers (e.g., R or Python), not much data is currently out in the open. In the particular case of the real estate market, to our knowledge, there are few open micro-data sets of housing markets available (Song et al., 2021).

To overcome these limitations, this paper presents a sort description of an open micro-data set of geo-referenced dwelling listings. 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 open geo-referenced micro-data set of the housing market in Spain. Moreover, the data set has been supplied directly by Idealista, and therefore is clean and free of download errors. 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. The data set is distributed as an R package, named 'idealista18', which can be accessed from the Github repository.

Data description

The open data set 'idealista18' is an R package composed of nine 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. The following subsections describe each object. A full description of the data is available in the help section of the package.

Dwelling listings

The dwelling listing of each city includes a set of characteristics for each dwelling published on the idealista real estate website as an ad. The dwelling listing has been included in the 'idealista18' package as an sf object (Pebesma, 2018). The name of

*Idealista is the major real estate listing website in Spain, and present in other southern european countries as Italy and Portugal the sf object containing the dwelling listing includes the name of the city, followed by '_Sale' (e.g., Madrid_Sale) and includes a total of 42 variables. Each sf object includes the complete set of listings corresponding to the four quarters of the year 2018. Table 1 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	Fourth	Total ads
Barcelona	17826	7951	12375	23334	61486
Madrid	21920	12652	15973	44270	94815
Valencia	9305	4655	5644	14018	33622

Table 1. 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 location of the dwelling (see Section Anonymizing the data set). Table 2 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 2, 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 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 2 show the mean values of these variables.

In addition to the variables listed in Table 2, the sf object includes a set of dummy variables with information about the basic characteristics of the dwelling. Table 3 shows the more relevant variables included in the sf object.

Neighboorhood polygons

The second block of data included in the 'idealista18' R package is the spatial features of the three cities divided into neighborhoods. There is an sf object for each city with the name of the city and the suffix '_Polygons'. Figure 1 shows the quantile maps of the

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 2. 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 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.

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 3. 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

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

[†] 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.

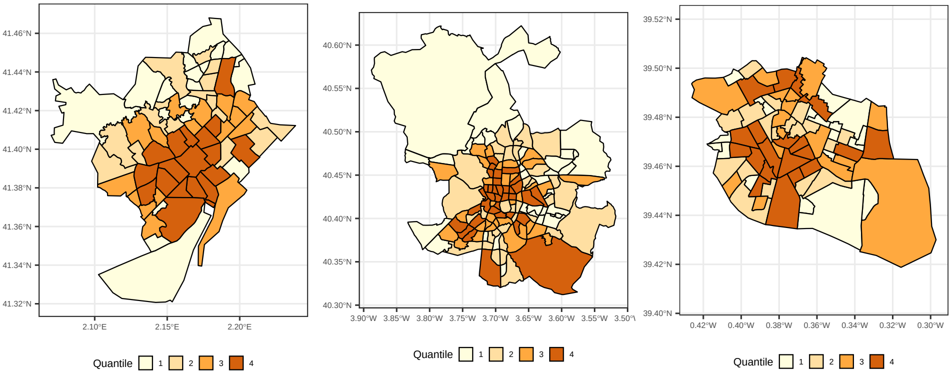

Figure 1. Quantile maps of the number of dwellings in each neighborhood. Boundary for Barcelona (Left), Madrid (Center), and Valencia (Right).

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.

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).

Points of Interest

The last block of data 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.

Anonymizing the data set

To comply with Spanish regulations, two variables were slightly modified to provide 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%. Since asking prices are usually multiples of 1,000, after the first price modification, the prices were aligned to multiples of 1,000.

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 $\alpha$ 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 $\alpha$
    until this stop condition;
    set $(X', Y')$ as the new location for the listing L
end
```

Algorithm 1: Coordinate displacement process for anonymisation purposes

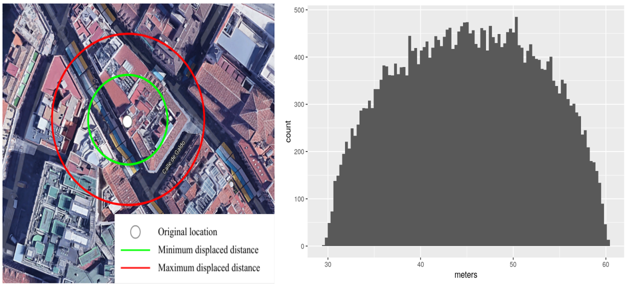

Figure 2. (Left) Masking coordinates. Spatial range. (Right) Coordinate displacement in meters (Valencia)

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 housing market and housing prices. Researchers can apply hedonic models with spatial effects, identifying housing submarkets or machine learning techniques. The data product can also be used for educational proposes and teaching activities.

References

Arbia G and Nardelli V (2020) On spatial lag models estimated using crowdsourcing, web-scraping or other unconventionally collected data. *arXiv preprint arXiv:2010.05287* .

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: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: 10.1007/s10109-021-00363-5.

Gupta A, Van Nieuwerburgh S and Kontokosta C (2022) Take the q train: Value capture of public infrastructure projects. *Journal of Urban Economics* : 103422.

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: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: 10.1111/pirs.12063.

Páez A (2021) Open spatial sciences: an introduction. *Journal of Geographical Systems* 23(4): 467–476. DOI: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:10.32614/RJ-2018-009.

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

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:10.1016/j.dib.2021.106877.