Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A geo-referenced micro-data set of real estate listings for the three largest Spanish cities from the Idealista website

Author One[1], Author Two[1]

ARTICLE INFO

Keywords:

Hedonic price analysis
Spain
Spatial econometrics
Machine learning
Geo-spatial analysis

ABSTRACT

This data article shares an open data product with big geo-referenced micro-data sets of 2018 real estate listings in Spain. The observations are for the three largest cities in the country: Madrid (n = 94,815 observations), Barcelona (n = 61,486 observations) and Valencia (n = 33,622 observations). These data were originally published on `idealista.com` real estate website. The data sets include the coordinates of properties (latitude and longitude), asking prices of each listd dwelling, and several variables of indoor characteristics. The listings were enriched with official information (building year of construction and built quality materials grade) plus other relevant geographical features such as distance to urban points of interest. Along with real estate listings, the data product also includes neighborhood boundaries for each city. These data sets are companions to a paper that examines the potential of machine learning techniques for spatial market segmentation in hedonic price analysis. The data product is offered in the form of a fully documented ‘R‘ package. This open data product is available for scientific and educational purposes, in particular for geo-spatial studies.

article reference to be inserted

Specifications Table

Every section of this table is mandatory. Please enter information in the right-hand column and remove all the instructions

Subject	Geography, Economics
Specific subject area	Spatial analysis, machine learning, hedonic price analysis

http://dx.doi.org/10.1016/j.dib.xxxx.xx.xxx

Type of data How data were acquired	Tables: dataframes and ‘sf‘ objects in ‘R‘ format Dwelling listing records provided by idealista.com [1] Spanish central cadastral registry [2] Open street map [3]
Data format	‘R‘ package, named {idealista18}
Parameters for data collection	Data has been directly downloaded from the sources, cadastral and idealista website data has been merged based on geographical location for each record.
Description of data collection	idealista provided the complete record set cadastral information has been downloaded the open records published quarterly open street map has been downloaded from its open API
Data source location	Institution: Idealista City/Town/Region: Madrid, Barcelona, Valencia Country: Spain Latitude and longitude samples/data: EPSG:4326
Data accessibility	Repository name: GitHub Direct URL to data: https://github.com/paezha/idealista18
Related research article	Author One et al., Using machine learning to identify spatial market segments: A reproducible study of major Spanish markets, Comput Environ Urban Syst. In Press.

Value of the Data

- A cleaned and enriched dataset consisting of real estate listings for three major cities in Spain. It has been constructed to analyze the impact of using machine learning models to identify spatial market segments when building house price hedonic models.
- The dataset can be used to extend the topic of automatic or semi-automatic identification of house market segments.
- The neighborhood boundaries combined with spatial patterns can be used to analyse the suitability of these boundaries as spatial dummy variables for real estate analyses purposes.
- The dataset can be enlarged with complementary spatial information to develop hedonic models.
- The data can be processed by quantitative analysis and statistical modeling to study the different factors that affect house prices in the three locations.
- Identification of spatial patterns in the real estate scope using the geo-referenced data points. For either value or urban patterns discovery.

Data Description

This open data product [4] is composed of several data objects corresponding to three major Spanish cities: dwelling listings, neighboorhood polygons and a set of Points of Interest (POI) for each city[1]. The data set is distributed in the form of an 'R' package, named {idealista18}. All spatial objects such as polygons and points are distributed as simple features objects (class 'sf' in 'R'). Spatial objects include geodetic coordinates using the *EPSG:4326* coordinate reference system.

The first block of data integrates properties published on idealista real state website [1]; each file contains the complete set of listings for each of the three cities in the package, corresponding to the four quarters of year 2018. Each record contains the key found in the listing ad[2] plus a number of additional attribues from the Spanish cadastre [2]. Cadastral information is described in Table 2, including the the prefix *CAD* in the variable name. Cadastral features asssignment is done by assigning the features of the nearest parcel to the coordinates *LATITUDE* and *LONGITUDE*. The measure scales for each variable are defined according to the theoretical framework proposed by [5] that defines four scales: nominal, ordinal, interval and ratio.

Table 2: Description of the variables in the listing data sets

Variable	**Mesurement scale**	**Description**
ASSETID	Identifier	Unique identifier of the advertisement
PERIOD	Nominal (Date)	Expressed as YYYYMM, indicates the quarter when the ad was extracted. We used YYYY03 for the 1st quarter, YYYY06 the 2nd, YYYY09 for the 3rd and YYYY12 for the 4th
PRICE	Interval	Asking price for the ad at idealista expressed in euros
UNITPRICE	Interval	Asking price in euros per square meter (constructed area)
ROOMNUMBER	Ordinal	Number of bedrooms
BATHNUMBER	Ordinal	Number of bathrooms
HASTERRACE	Nominal	Dummy variable for terrace (takes 1 if there is a terrace, 0 otherwise
HASLIFT	Nominal	Dummy variable for lift (takes 1 if there is a lift in the building, 0 otherwise)
HASAIRCONDITIONING	Nominal	Dummy variable for air conditioner (takes 1 if there is air conditioner, 0 otherwise)
AMENITYID	Nominal	Indicates the amenities included (1 - no furniture, no kitchen amenities, 2 - kitchen amenities, no furniture, 3 - kitchen amenities, furniture)
HASPARKINGSPACE	Nominal	Dummy variable for parking (takes 1 if parking is included in the Ad, 0 otherwise)
ISPARKINGSPACEINCLUDEDINPRICE	Nominal	Dummy variable for parking (takes 1 if parking is included in the Ad, 0 otherwise)
PARKINGSPACEPRICE	Interval	Asking price of parking space in euros
HASNORTHORIENTATION	Nominal	Dummy variable for orientation (takes 1 if orientation is North in the Ad, 0 otherwise) - Important note: orientation features are not orthogonal features, a house oriented to the north can be also oriented to the east
HASSOUTHORIENTATION	Nominal	Dummy variable for orientation (takes 1 if orientation is South in the Ad, 0 otherwise) - Important note: orientation features are not orthogonal features, a house oriented to the north can be also oriented to the east

[1] The data has been provided by Idealista, the major real estate listing website in Spain, and present in other southern european countries as Italy and Portugal.

[2] The term *ad* or *listing* is used interchangeably to refer to a property advertised on the website

HASEASTORIENTATION	Nominal	Dummy variable for orientation (takes 1 if orientation is East in the Ad, 0 otherwise) - Important note: orientation features are not orthogonal features, a house oriented to the north can be also oriented to the east
HASWESTORIENTATION	Nominal	Dummy variable for orientation (takes 1 if orientation is West in the Ad, 0 otherwise) - Important note: orientation features are not orthogonal features, a house oriented to the north can be also oriented to the east
HASBOXROOM	Nominal	Dummy variable for boxroom (takes 1 if boxroom is included in the Ad, 0 otherwise)
HASWARDROBE	Nominal	Dummy variable for wardrobe (takes 1 whether the property has wardrobes, 0 otherwise)
HASSWIMMINGPOOL	Nominal	Dummy variable for swimming pool (takes 1 if swimming pool is included in the Ad, 0 otherwise)
HASDOORMAN	Nominal	Dummy variable for doorman (takes 1 if there is a doorman in the building, 0 otherwise)
HASGARDEN	Nominal	Dummy variable for garden (takes 1 if there is a garden in the building, 0 otherwise)
ISDUPLEX	Nominal	Dummy variable for bachelor apartment (referred as studio in Spain) (takes 1 if it is a bachelor apartment, 0 otherwise)
ISINTOPFLOOR	Nominal	Dummy variable indicating if the apartment is located in the top floor (takes 1 on the top floor 0 otherwise)
CONSTRUCTIONYEAR	Interval	Construction year (source: advertiser)
FLOORCLEAN	Ordinal	Indicates flat floornumber starting from the 0 value for ground floor (source: advertiser)
FLATLOCATIONID	Nominal	Indicates the kind of views the flat has (1 - external, 2 - internal)
CADCONSTRUCTIONYEAR	Interval	Construction year as of cadastral source (source: cadastre), note this figure can differ from the one given by the advertiser
CADMAXBUILDINGFLOOR	Ordinal	Max building floor (source: cadastre)
CADDWELLINGCOUNT	Interval	Dwelling count in the building (source: cadastre)
CADASTRALQUALITYID	Ordinal	Cadastral quality (source: cadastre). 0 Best - 10 Worst
BUILTTYPEID_1	Nominal	Dummy value for flat condition: 1 new development and 0 otherwise
BUILTTYPEID_2	Nominal	Dummy value for flat condition: 1 second hand to be restored 0 otherwise (*source: advertiser*)
BUILTTYPEID_3	Nominal	Dummy value for flat condition: 1 second hand in good condition 0 otherwise (*source: advertiser*)
LONGITUDE	Interval	Longitude, geographical coordinate
LATITUDE	Interval	Latitude geographical coordinate
geometry	Geometry	Geometry for the elements. A point with X, Y coordinates

In addition to the information shown in Table 2, distances from each dwelling to Point of Interest are included. Table 3 lists these variables for each city.

Table 3: Description of variables of neighborhood polygons data set

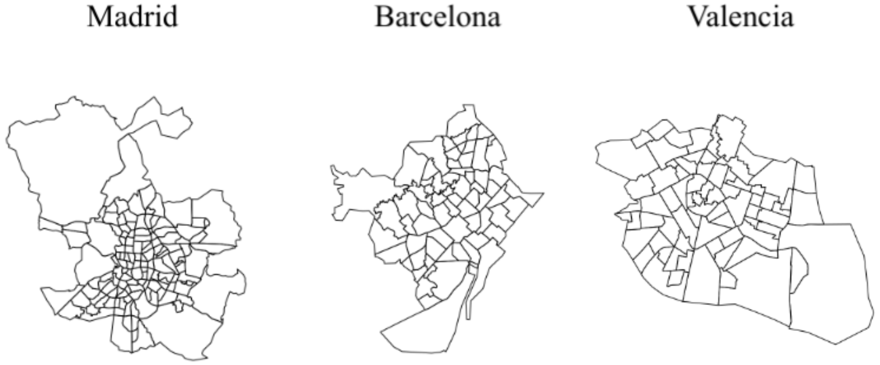
City	**Variable**	**Mesurement scale**	**Description**
Madrid	DISTANCE_TO_CITY_CENTER	Interval	Distance in Km to the city center (Puerta del Sol)
	DISTANCE_TO_METRO	Interval	Distance in Km to the nearest subway station

	DISTANCE_TO_CASTELLANA	Interval	Distance in Km to the Paseo de la Castellana Street
Valencia	DISTANCE_TO_CITY_CENTER	Interval	Distance in Km to the city center (Plaza del Ayuntamiento)
	DISTANCE_TO_METRO	Interval	Distance in Km to the nearest subway station
	DISTANCE_TO_BLASCO	Interval	Distance in Km to the Blasco Ibáñez Avenue
Barcelona	DISTANCE_TO_CITY_CENTER	Interval	Distance in Km to the city center (Plaza de España)
	DISTANCE_TO_METRO	Interval	Distance in Km to the nearest subway station
	DISTANCE_TO_DIAGONAL	Interval	Distance in Km to the Diagonal Avenue

The record counts for each city in 2018 are: 94,815 listings for Madrid, 61,486 for Barcelona and 33,622 for Valencia. Note that the same listing can be found in more than one period when a property listed for sale in one quarter was sold in a subsequent quarter.

The second block of data include the spatial features of the three cities divided in neighborhoods. Figure 1 shows the different neighborhoods for the three cities. The boundaries are based on the official boundaries but slighly adapted by idealista [3]. In practical terms we can assume they are the same, since the website simply collapses areas when they are sufficiently small in terms of number of ads. In the case of Madrid they just collapse four areas into two new ones.

Fig. 1. Neighborhood boundaries for Madrid, Barcelona and Valencia



There are a total of 73 neigborhoods in Barcelona, 135 in Madrid and 73 in Valencia. Each neighborhood has also two additional variables described in the Table 4.

[3] The criterion used to adapt this division is double, if an area is small enough and similar enough to another they merge both areas, on the other hand if the official area is not homogeneous it is then divided in a series of new polygons

Table 4: Additional variables for each city

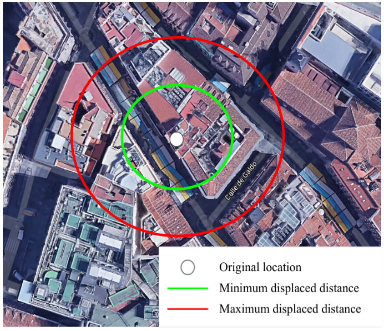
Variable	Mesurement scale	Description
LOCATIONID	nominal	Unique identifier for the neighborhood
LOCATIONNAME	nominal	Neighborhood name

The last block of data included in the data package is a set of Point of Interest of each city in ‘R‘ list format. These lists include three elements: (i) the coordinates of the city center (see Table 4 that identify the central business district); (ii) a set of points that define the main street of each city; and (iii) the coordinates of metro stations.

Experimental Design, Materials and Methods

As noted above, the data package contains the complete offering in three major Spanish cities for each of the four quarters of 2018, as drawn from idealista web site [1]. To comply with Spanish regulations, the listings are slightly modified to preserve their anonimity. The masking process consists of two steps, as follows: first the prices are obfuscated with the addition or substraction of a random percentage of their original values ranging from -2.5% to +2.5%. Since asking prices are not normally a completely continuous variable (sale prices are usually multiples of 1000 and rent prices are of 10), after the first price modification we finally align prices to multiples of 1000. Secondly, a spatial masking process was also implemented with the intention of keeping spatial properties of the original data set. The coordinates of each listing were displaced using a stochastic procedure. Effectively, the listings were recoded using coordinates contained in a maximm and minimum displacement circles, as shown in Figure 2. To preserve membership in a neighborhood, the spatial masking procedure was constrained to ensure that the masked coordinates are in the original neighborhood of the listing.

Fig. 2. Masking coordinates spatial range. Source: own elaboration



The 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 in a different neighborhood. This ensures that neighborhood attributes are preserved.

Figure 3 shows the histogram of displacements in meters for all listings in the city of Valencia; the average distance between the original and masked coordinates is 45 meters.

```
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

Fig. 3. Coordinate displacement in meters Valencia. Source: own elaboration

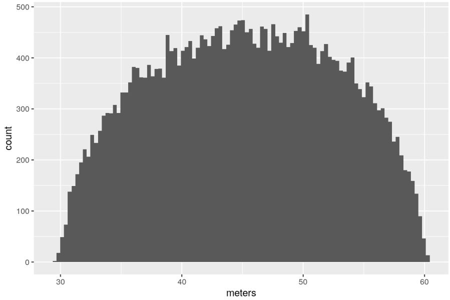


Figure 4 shows the spatial distribution of the original records compared to spatial distribution after masking.

Acknowledgments

The authors wish to thank X for their support in the paper revision and Y for collecting and cleaning the spatial data.

Declaration of Competing Interest

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

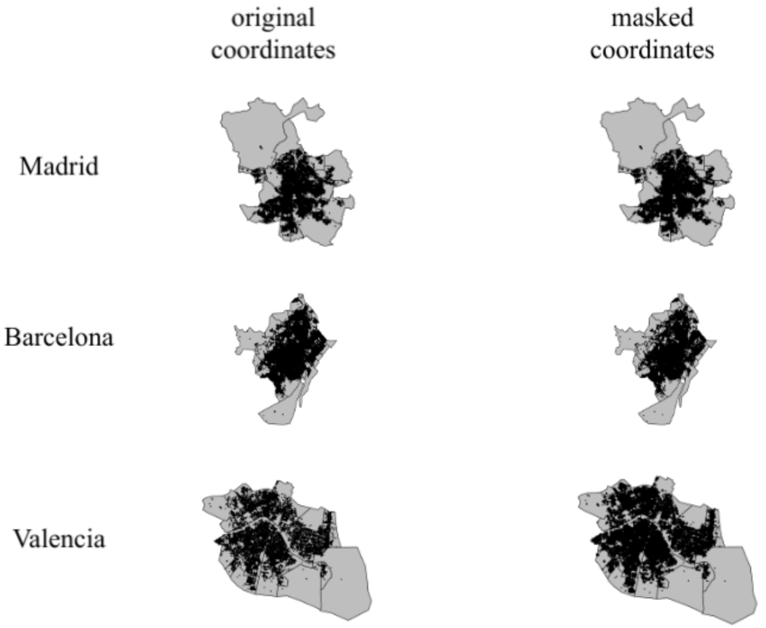
References

[1] idealista, http://www.idealista.com , http://www.idealista.com, 2018.

[2] Registro Central del Catastro, https://www.sedecatastro.gob.es/ , 2021.

[3] OpenStreetMap contributors, Planet dump retrieved from https://planet.osm.org , https://www.openstreetmap.org, 2021.

Fig. 4. Spatial distribution of ads (before and after masking). Source: own elaboration



[4] D. Arribas-Bel, M. Green, F. Rowe, A. Singleton, Open data products-a framework for creating valuable analysis ready data, Journal of Geographical Systems 23 (2021) 497–514.

[5] S. S. Stevens, et al., On the theory of scales of measurement (1946).