Contents lists available at ScienceDirect

Data in Brief

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

Data Article

idealista18: A data package with real estate information in three major Spanish markets from the Idealista database

David Rey Blanco[a], Pelayo González Arbues[a], Fernando López Hernández[b], Antonio Páez[c,*]

[a] *idealista, Plaza de las Cortes 5, 28014 Madrid, Spain*
[b] *Facultad de CC de la Empresa, C/ Real, 3. 30201 Cartagena, Murcia (Spain)*
[c] *School of Earth, Environment and Society, McMaster University, 1280 Main St W, Hamilton, Ontario L8S 4K1 Canada*

ARTICLE INFO

Keywords:

Property values
Spain
Spatial analysis
Machine learning
Hedonic price analysis

ABSTRACT

This dataset contains three items for each of the three major cities in Spain: Madrid, Barcelona and Valencia. The first data set contains real estate listings published on idealista portal in 2018. All listings have been enriched with cadastral information (i.e. building year of construction, built quality materials grade) plus some geographical features such as distance to relevant city areas and the coordinates themselves. To comply with european personal protection laws, we have processeed some sensitive variables yet preserving their spatial properties. The second data sets contains the neighborhood boundaries for each city, and third sets comprise a set of key points of interest for each municipality. This dataset is suitable to house market analysis, hedonic house price models and other spatial research related with real estate markets.

¡falta cita/referencia al artículo¿

Specifications Table

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

*Corresponding author: Tel.: +1-905-525-9140 ext 26099
e-mail: paezha@mcmaster.ca (Antonio Páez)

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

Subject	Geography, Economics
Specific subject area	Spatial analysis, machine learning, hedonic price analysis
Type of data How data were acquired	Tables multi-family listing records given by idealista portal [1] Spanish central cadastral registry [2] Open street map [3]
Data format	Spatially masked
Parameters for data collection	Data has been directly downloaded from the sources, cadastral and idealista 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	D. Rey Blanco, P. González Arbues, F. López Hernández, A. Páez, 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

The data sets is composed by three items for three major spanish cities: quarterly single family listings, neighboorhood polygons and a set key of Points of Interest for each city. All spatial features, such as polygons and points, are expressed in geodetic coordinates using the *EPSG:4326* coordinate reference system. The first block of data integrates properties published on idealista web site [1]; each file contains the complete offering for a city for the four quarters

in 2018. Idealista is the major real estate listing portal in Spain and also present in other southern european countries as Italy and Portugal. Each record contains the key found in listing ad[1] plus a number of additional attribues from the Spanish cadastre [2]. The latter are described in the table 2, and the names for all these variables start with the prefix *CAD*. 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 has been defined according the theoretical framework proposed by [4] that defines four scales: nominal, ordinal, interval and ratio.

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

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)
ADTYPOLOGYID	Nominal	Residential building type: multi-family: *home*, single-familiy: *chalet*
ADOPERATIONID	Nominal	Operation type for the ad: *sale* or *rent*
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 AA (takes 1 if there is a AA, 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	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
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

[1] we will use the term *ad* or *listing* interchangeably to refer to a house published on the portal

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*)
DISTANCE_TO_CITY_CENTER	Interval	Distance to center of city in Km
geometry	Geometry	Geometry for the elements. A point with X, Y coordinates

In addition to the common features each city will have a set of spatial additional features, in particular refered to the distance to a major street.

Table 3: Description of variables of neighborhood polygons data set

City	Variable	Mesurement scale	Description
Madrid	DISTANCE_TO_CASTELLANA	Interval	Distance in km to the Paseo de la Castellana Street
Valencia	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_DIAGONAL	Interval	Distance in km to the Diagonal Avenue

The record count for each city is: 156,016 listings for Madrid and 84,280 for Barcelona and 79,360 for Valencia. It is important to note that the same listing can be found in more than one

period, what means that a house was uploaded for sale in one quarter but was sold in a subsequent quarter.

The second part contains the polygons for the different neighborhoods for such cities as we can see in the figure 1. This boundaries are based on the official boundaries but slighly adapted by idealista [2]. In practical terms we can assume they are the same, since the website simply collapsed those areas if they are small enough in terms of number of ads. In the case of Madrid they just collapse four areas in two new ones.

Fig. 1. Neighborhood boundaries for Madrid, Barcelona and Valencia. Source: own elaboration

We have a total of 73 neigborhoods for Barcelona, 135 for Madrid and 73 for Valencia. Each neighborhood has also two additional variables described in the table 4.

Table 4: Additional variables for each city

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

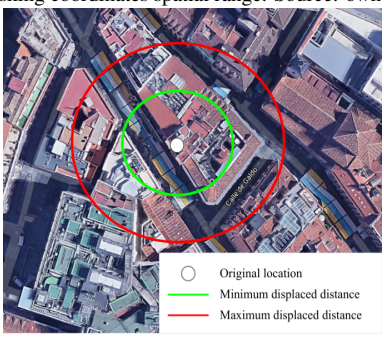
Experimental Design, Materials and Methods

The data encompassed three major cities in Spain: Madrid, Barcelona and Valencia, for each municipality we provide listing prices for multi-familiy homes on a quarterly basis. The files contain the complete offering for each of the four quarters in 2018 on idealista web site [1]. Idealista is the major real estate listing portal in Spain and also present in other southern european countries as Italy and Portugal.Given the spanish regulatory restrictions idealista listings are slighly anonymized, in order to agree regulatory requirements guaranteeing their attributes and spatial properties. This process take two steps, the first consists of the obfuscation of prices adding or substracting a random percentage of their original values ranging from -2.5% to +2.5%.

[2]the criterium 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

Since asking prices are not normally a completely continuous magnitude (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.Finally we carry out a spatial masking process that intends to keep spatial properties of the original data set. As we can see in the figure 2 the area where the new point (masked) would be located will be taken randomly within maximum and minimum displacement distances circles. In order to preserve the nature of the neighborhood the house we make sure the new point would fall in the original neighborhood, otherwise we look for a new masked place.

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



The algorithm 1 iteratively displace the coordinates of each listing with a minimum distance and a maximum distance with the restriction that every new location must fall in within the original neighborbood 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 anonymisation purposes

In the figure 3 we display the histogram with the displacement in meters for all listings in the city of Valencia, as average the displacement is 45 meters.

In the figure 4 we can see the spatial distribution of the original records compared to spatial distribution after masking.

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

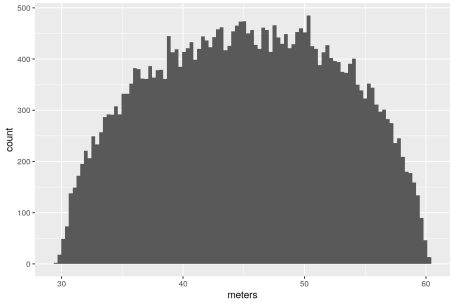
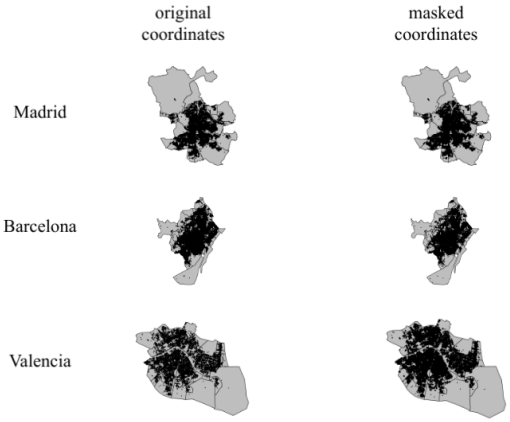


Fig. 4. Spatial distribution of ads (before and after masking). Source: own elaboration



Acknowledgments

The authors thank Alessandro Galesi for his support in the paper revision and Juan Ramon Selva for collecting and cleaning the spatial data.

Declaration of Competing Interest

The authors declare that they have no 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.
[4] S. S. Stevens, et al., On the theory of scales of measurement (1946).