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

Referee: 1

Comments to the Author

Dear Authors, Thank you for the opportunity to review this paper. I would appreciate it if you would highlight the changed text in the next submission. Otherwise it is difficult to identify which parts have changed. The unknown quality of this one year randomised data is the main issue for potential users and I repeat my suggestion that this data should be validated for the benefit potential readers and users. Below is a list of comments and questions which need to be addressed.

Thank you for your constructive comments. Below we describe how we improved the paper in response.

---

1. Please clearly state the nature of the data in the abstract. It is worthwhile to mention the relocated and randomised dataset in the abstract. This house price dataset only has 189,923 records, which could be not really be described as a big dataset as you describe it in the abstract. Except the authors could provide an evidence/reference on why this is big dataset.

Response:

We appreciate this comment. There are a few publicly available datasets for hedonic price analysis, but they tend to be smaller. For example, Bonifaci and Copiello (2015) includes $n = 1,042$ observations for Padua, in Italy; Del Giudice, De Paola, and Forte (2018) share a dataset with $n = 576$ observations relating to rental prices in Naples, Italy; Solano Sánchez et al. (2019) present a dataset with $n = 1,623$ daily rental prices in Seville, Spain; and the dataset of Fuerst and Haddad (2020) includes $n = 4,201$ property prices geocoded to the level of nine regions in England and Wales. 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 ($n = 25,984$). In response to your comment we drop the terminology of "big data" and describe our dataset as "large" instead.

---

2. Please check the typographical errors before your next submission. For example, the "thirdr" in your table1.

Response:

Thank you for highlighting this. We have carefully proof-read the document for typos and other errors.

---

3. Please use part of your answer to my previous fourth question in the manuscript. Your answer is quite useful for your future users, especially those researchers who are not familiar with house price data in Spain. I quote your previous answer below:

"idealista covers fairly well all segments in the Spanish market, both individual and professional advertisers. This dataset comprehends information about listing prices, thus it contains the market situation from the offer. Nevertheless, actual sale prices information is not publicly available as it is information that 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 using the

offer perspective, they keep strong correlation between idealista and transaction prices can be established see Banco de España. Even though a great extent of correlation does exist between official transactional and asking records, both databases can be taken as complementary and are of great interest when studying asking-transaction price gaps or the relation between listing site demand variables (ie. ad contacts or ad views) and price gaps. The main real estate portals in Spain are idealista and Fotocasa, further away are Habitaclia and Milanuncios, the latter focusing almost exclusively on private individuals. In September 2021, according to data from Similarweb (a site specialized in sites' traffic volume comparison), in Spain, there were a total of 103 million page views on real estate portals, out of page views on real estate portals, of which the four main portals, idealista, Fotocasa, Habitaclia, and Pisos.com, in that order, accounted for 94% of the traffic. Another relevant feature of this sector is that it is highly concentrated, idealista being the leader by far with 58.6 monthly million visits (57% of the total traffic) versus its immediate competitor with 19.9 million visits (19.3 % of total traffic). In terms of the evolution over time of interest in the content of each portal, based on data from Google Trends, the leadership of idealista Trends, the leadership of idealist"

Response:

Thank you for this suggestion. We have adapted this answer to expand the text in the section "Data Description".

---

4. For data description, it is worth showing solid evidence of the quality of this dataset. For example, what is the data coverage of this data compared to all the residential properties in the city? Listing the advertised volume is not very helpful for users who wish to understand the quality.

Response:

To better place in context the coverage of the {idealista18} dataset, Table 1 now 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[1] for 2018 also shows that the number of transactions in 2018 are equivalent to 81.3% of idealista listings in Barcelona, 80.8% in Madrid, and 91.1% in Valencia. While it is not possible to link actual transactions to idealista listings, this gives a quantitative sense of the coverage of the data package.

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

---

5. Given the asking price is a resized version of the original asking price with a random percentage between -2.5% and +2.5%, please elaborate on how this influences your quantitative research results and findings. Please provide critical evidence to explain how this randomised price will not influence academic research. For example, how does the house price variation at local level change after the price is randomised. To what degree will this influence a typical

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

hedonic house price research approach? I am concerned about whether this randomised data could bias conclusions.

Response:

We appreciate the opportunity to delve more deeply into this issue, as it may be of interest to users of the data package. The short version of this story is that adding a small amount of random noise to the prices ($\pm 2.5\%$) does not introduce bias but does increase the variance of the masked variable. This can be shown with some standard algebra of random variables.

Now the longer story.

The masked prices are given by:

$$P = RP \cdot \epsilon$$

where ϵ is 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 2 reports the mean and the standard deviation of the prices in the package, and the standard deviation of the raw prices calculated using the formula above. 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).

Table 2: 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)

6. In figure 1, is it possible to standardise this data rather than showing only the number of dwelling records in your house price data? I guess the city centre always has high advertised volumes but it also has a high density of dwellings. It is worthwhile to show values standardised by the number of dwellings.

Response:

After this comment we included a new figure in the paper to display the percentage of listings relative to the total number of dwellings by district (see Figure 1 below).

Thank you again for your generous feedback. We hope that our responses will prove satisfactory and you will agree that the paper makes a worthwhile contribution to growing catalog of datasets to support research on real estate markets.

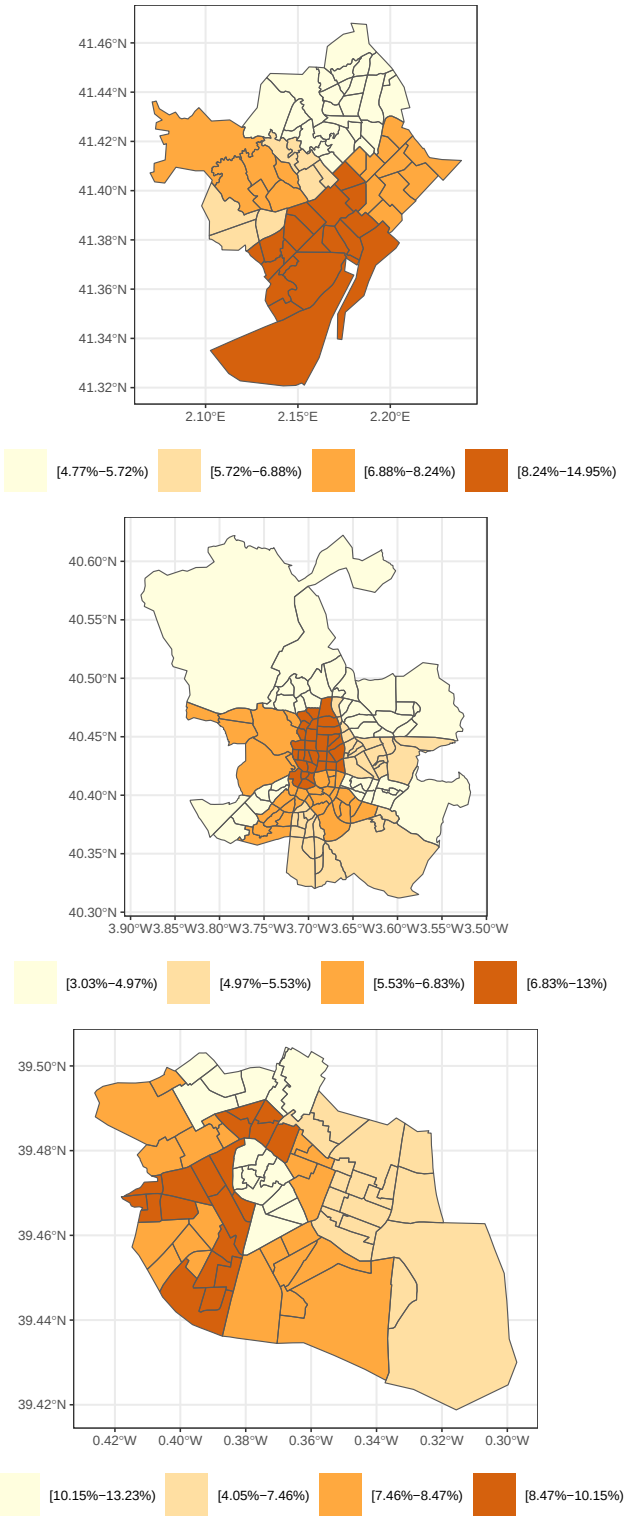

Figure 1: Percentage of listings relative to total number of dwellings by district. Boundary for Barcelona (Top), Madrid (Center), and Valencia (Bottom).

References

Bonifaci, Pietro, and Sergio Copiello. 2015. "Real Estate Market and Building Energy Performance: Data for a Mass Appraisal Approach." Journal Article. *Data in Brief* 5: 1060–65. https://doi.org/https://doi.org/10.1016/j.dib.2015.11.027.

Del Giudice, Vincenzo, Pierfrancesco De Paola, and Fabiana Forte. 2018. "Housing Rental Prices: Data from a Central Urban Area of Naples (Italy)." Journal Article. *Data in Brief* 18: 983–87. https://doi.org/10.1016/j.dib.2018.03.121.

Fuerst, Franz, and Michel Ferreira Cardia Haddad. 2020. "Real Estate Data to Analyse the Relationship Between Property Prices, Sustainability Levels and Socio-Economic Indicators." Journal Article. *Data in Brief* 33: 106359. https://doi.org/10.1016/j.dib.2020.106359.

Solano Sánchez, Miguel Ángel, Julia Margarita Núñez Tabales, José María Caridad y Ocerin, José Antonio C. Santos, and Margarida Custódio Santos. 2019. "Dataset for Holiday Rentals' Daily Rate Pricing in a Cultural Tourism Destination." Journal Article. *Data in Brief* 27: 104697. https://doi.org/10.1016/j.dib.2019.104697.

Song, Yena, Kwangwon Ahn, Sihyun An, and Hanwool Jang. 2021. "Hedonic Dataset of the Metropolitan Housing Market – Cases in South Korea." Journal Article. *Data in Brief* 35: 106877. https://doi.org/10.1016/j.dib.2021.106877.