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

Dear Authors, Thank you for the opportunity to review your paper. This manuscript is about the publishing of a one-year relocated and randomised listing price data set for three cities in Spain. Both the asking price and coordinates information are anonymised which limits the usefulness of the data for teaching and academic research. Given the house price data is partly from the Idealista company, it is not clear under which kind of licence the data is being issued and to what extent it may be considered open. The unknown quality of this one year randomised data is the main issue for potential users. Having read the paper many times, I believe that the authors have created the idealista18 package (https://github.com/paezha/idealista18). The Idealista18 Github readme file is too simple to guidance users on how to access the data with the R package. So I am not sure it is the right one. The authors are to be commended for their intention to make open real estate data available in an R package which should render it easily accessible for teaching and research. Below is a list of comments and questions that I had which may need to be addressed in future editions of this work.

1. Please reference the package and dataset and include in the text and in the package the license under which the data is released. The Idealista18 github site is unclear on how to access the data with the package.

We really appreciate this suggestion. We have included the license under which the data is distributed in the main text of the introduction. We are also working to publish the dataset via the R package usign CRAN. In addition we have also created a microsite for the package that shows how to access the data.

2. Please state your anticipated research impacts from the publication of this data set in the introduction. It would be useful to state how this one-year relocated, and randomised house price data might be best used by academics in research and teaching. The relocated and randomised data may be difficult to further enrich with other variables, the authors need to show how the 44 variables is enough for future users.

Thanks for this comment. We have added a discussion about these concerns in the introduction.

3. The method of randomisation is described but it would be advantageous to publish the code for this so that researchers can evaluate the impact of this kind of randomisation on this and other data sets. Does the method, as seems likely, have a differential effect on data from high density areas as compared with low? To what extent are analyses likely to be confined to neighbourhood levels of resolution given the method of spatial randomisation? It would be useful to have some idea of the size of the neighbourhoods to facilitate comparison with other spatial data sets

The possible advese effects due to mislocation assets after coordinate randomisation is mitigated by the imposition of a boundary constraint in the relocation process. This constraint keeps the listing in the same neighborhood it belonged before relocation. For the process clarity sake we released the anonymization process in the R package code on github. microsite

4. More details on the data collection methods of Idealista would be useful in interpreting the data. Does Idealista concentrate on particular market segments or not? Are the list prices good indicators of actual sale prices? Are there sources from which actual paid prices could be identified?

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 idealista has been shared from 2004 to 2015, subsequently consolidating idealista as the leader by a wide margin.

5. It would be useful for the authors to evaluate the datasets in the light of the FAIR guiding principles for research data stewardship (findability, accessibility, interoperability, and reusability).

We discuss our efforts to comply with FAIR guiding principles in the data description of the revised paper.

6. Given the asking price has resized the original asking price with a random percentage between -2.5% and +2.5%, please elaborate on how this influences quantitative research results and findings with this data.

We are aware that performing data obfuscation in our dataset has a trade-off between data privacy and data utility. The data anonymization process maintains data utility and data properties for most empirical research as we are adding a white noise perturbation to the original prices and coordinates. This dataset is expected to be used to compare different valuation algorithms and in this sense, the random disturbances in the data should not cause major concerns.

However, for some applications researchers should proceed with care with this dataset. We would like to warn researchers that the spatial perturbation we introduced in the data might impact the computation of travel times between the properties' location and Point of Interest coordinates. In this case, travel times wouldn't be completely accurate although they are likely to be good enough for most accessibility measure computations. We must say in any case that we decided to take these perturbation since they are limited when it comes to walk mode travel times. As the average displacement length is around 30 meters, this distance would take an adult, on average, about 21-40 seconds at a normal pace Travel times walk.

7. Please kindly add in a usage notes section how this data can be useful for the EPB audience. For example, introduce how this anonymised house price data can be integrated with other data and research. The original data could be a nice data set for a series of housing related papers. If the authors are unable to publish this original asking price, it maybe worthwhile to publish the research based on this data rather than using the randomised data. House price is locationally sensitive, but both the location and price data in this manuscript are altered by anonymisation. The authors need to show for which analyses the data remains valid despite anonymisation.

We have added some examples of research topics that can be addressed with this dataset in the introduction.