Using machine learning to identify spatial market segments: A reproducible study of major Spanish markets - Response to reviewers

We would like to begin by thanking you for the careful, thoughtful review of the paper. We appreciate the constructive criticism and the opportunity to improve the paper. For your convenience, a version of the paper with all changes tracked is included in this revision.

First Comment

1.(a) The authors claim that machine learning techniques are often black boxes generating results which are often counterintuitive and estimates which are not stable enough to support, e.g., the decision making of commercial banks in their lending process by estimating a correct collateral value. Obviously, their approach avoids this issue by simply using classification tree technique in the first step. However, there are causalmachine learning approaches nowadays, which overcome such problems (at least to some extent) (see, e.g., Knaus, Lechner, and Strittmatter (2021)). I believe that the paper would benefit from such a discussion.

> This is a fair point. An emergent body of research aims at increasing the interpretability of machine learning methods, including Du, Liu, and Hu (2019), Murdoch et al. (2019), and most recently, as you point out, by Knaus, Lechner, and Strittmatter (2021). This is an area of research that is quickly evolving, although it is not without critics (e.g., Rudin 2019). Currently, existing approaches depend on fairly strong assumptions. For example, the causal forest framework (Wager and Athey 2018) assumes that the leaves of trees are sufficiently small so that paired treatment indicators and outcomes behave as if they were a randomized experiment. Assuming independence is often inappropriate in the analysis of spatial data, and techniques that correctly treat spatial dependencies are mature. It is possible that in the future interpretable machine learning techniques will address these issues, so we are advised to pay attention to this stream of research.
>
> We have added a footnote to expand our discussion about the potential of interpretable machine learning.

Second Comment

2.(a) On page 5, the authors highlight the advantage of their modeling strategy inidentifying market segments compared to the one proposed by Füss and Koller (2016). It is not entirely clear to me why the use of prices is a better choice. The use of residuals might be adequate as well if the model provides causal effects (see previous comment).In other words, what kind of information about the location are incorporated into the price which we do not see in the residuals. I believe that this essential part needs a more detailed discussion.

> Thank you for this comment. Regression tree estimates are the mean of the values contained in the volume of a leaf, which is to say a constant value. In a geographical application the leaves are mutually exclusive and collectively exhaustive partitions of geographical space. Using the residuals as a second step of the modelling strategy automatically results in spatial autocorrelation, since all properties in the same segment will be assigned estimated residuals that are identical. The issue here is that by inducing spatial autocorrelation in the second step some of the spatial information about location is obscured (i.e., there is zero spatial variation in the estimation for a given market segment). Our method avoids this by modelling the prices first to obtain market segments with

good homogeneity properties. Any spatial autocorrelation is dealt with by means of the spatial econometric model in the second step.

This is discussed in the text of the revised version of the paper.

Third Comment

3.(a) The empirical analysis only includes cross-sectional data for 2018Q4. However, it would be interesting to see how homogeneous neighborhoods change over time and space, e.g., due to gentrification.

Thank you for this fascinating suggestion. We are not aware of any research that investigates the stability of spatial market segments. A search on Web of Science using the following arguments returns 33 documents:

housing (Topic) AND space or spatial or geograph* (Topic) AND "market segments" or "market segment" (Topic)

A search using the following arguments in contrast returns zero documents:

housing (Topic) AND space or spatial or geograph* (Topic) AND "market segments" or "market segment" (Topic) AND season* or stability (Topic)

It is possible that spatial market segments experience seasonality and/or other temporal trends. Given the dearth of knowledge on this matter, we suggest that this question is an interesting topic for future research.

Out of curiosity, we modelled the market segments for each of four quarters covered by our data sets (see Figs. 1, 2, and 3 in this letter). As seen in the figures, the spatial market segments for 2018 are relatively stable for Barcelona and Madrid, but change somewhat more noticeably for Valencia in the course of the year. The differences need to be understood in the context of the coefficients estimated by the model for each segment, which may be still relativelly homogeneous. In any case, it is difficult to say how much of the variation observed could be random, or whether changes in the spatial segments cycle over longer periods of time in seasonal patterns (e.g., that the market segments change between winter and summer).

A limitation that we face to study more in depth the question of gentrification and the stability of market segments (in addition of course to space in the current paper), is the fact that gentrification and seasonality processes probably operate over longer periods than the year we investigate here. Furthermore, to publicly release this data sets we had to obtain permission from the legal department of idealista, and a practical constraint is that we were not allowed to share data sets spanning several years.

That said, our argument is that the two-step approach produces good results on a cotemporaneous test sample. This suggest that the method is useful in nowcasting or short-term forecasting. We have added a caveat in the conclusings about the potential seasonality of the market segments, so any attempt to use them for longer term forecasts should be treated with caution, as follows:

"One direction for future research is to investigate the temporal stability of spatial market segments. It is well known that there are seasonal effects in housing markets, but an open research question is whether spatial market segments experience seasonal variations, both in terms of their geographical extent as well as the magnitude of their effects. Another possibility is that there are longer term trends (e.g., gentrification) that could affect the spatial configuration of the market segments. Both seasonality and/or longer term trends would require multi-year data sets, compared to the single-year data set that we used for this research. For the time being, it is important to note that the results presented in this paper support the argument that the two-step method described in this paper performs well for now-casting or relatively short term forecasts. Given the dearth of information about seasonality and temporal stability of spatial market segments, any attempt to use them for longer term forecasts should be done with caution."

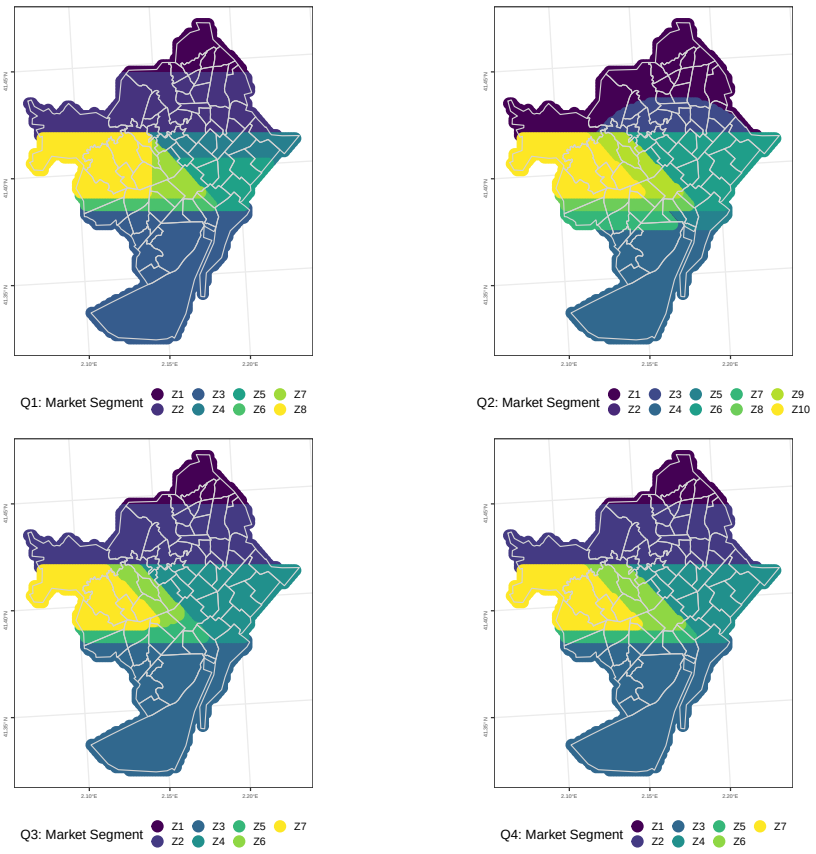

Figure 1: Spatial market segments according to Stage 1 classification tree for Barcelona Q1 through Q4

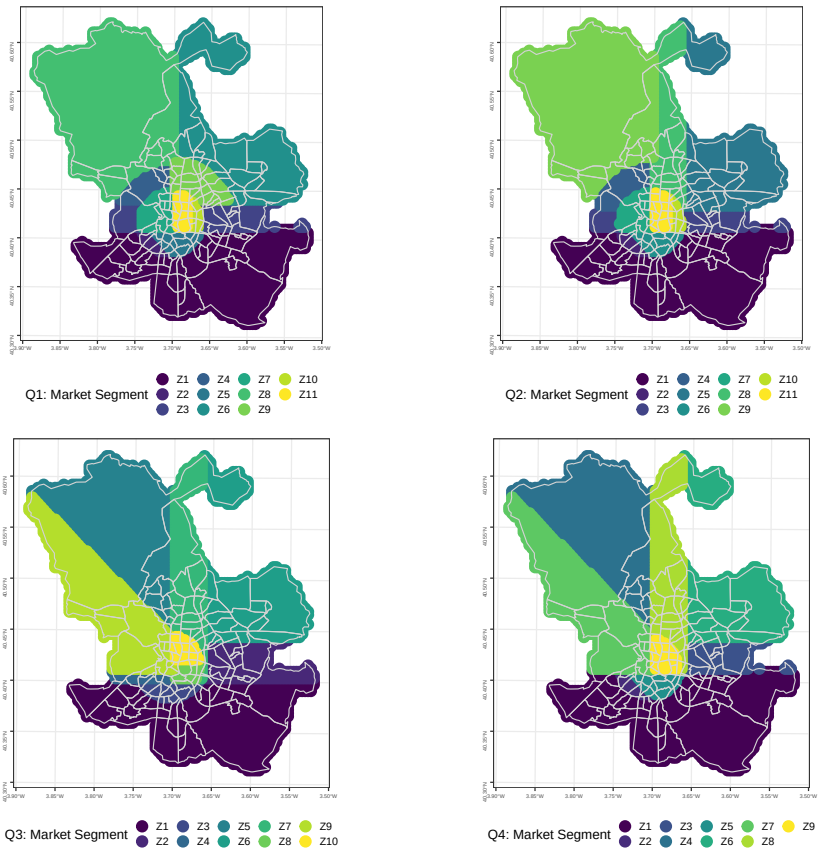

Figure 2: Spatial market segments according to Stage 1 classification tree for Madrid Q1 through Q4

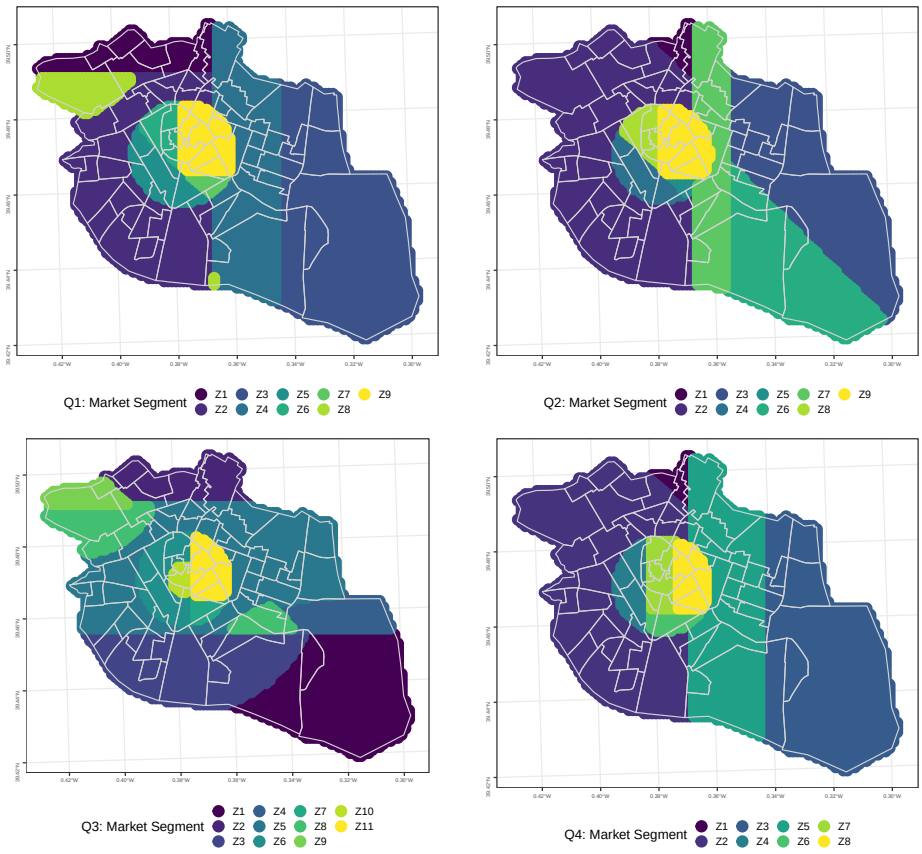

Figure 3: Spatial market segments according to Stage 1 classification tree for Valencia Q1 through Q4

3.(b) I also do not understand why the authors include the three cities which might by nature be more homogenous than smaller cities or rural areas (or high versus less populated areas). How do they differ in terms of city size, average income, local gdp etc.?

> In regard to the selection of the cities: at the beginning, we thought to release data just on Madrid but we believed that releasing open data for the three largest cities in Spain would be attractive for other researchers and that it would allow us to test the consistency of results in two big metropolises and a medium-size city such as Valencia. Our choice of cities was determined purely by the permissions we could obtain to share the data publicly to increase the reproducibility of the research. A systematic comparison of conditions by city attributes is beyond the scope of the present paper.

3.(c) It seems that some of the coefficients change sign among the cities. For instance, why does the price decreases when the number of rooms increases in Madrid?

> Thank you for this question. It is important to note that in models with spatial autocorrelation the coefficients are not the marginal effects. Due to the ripple effect of the spatial lag (multiple variables enter the derivatives), there are indirect effects as well. The sum of the direct and indirect effects is the total marginal effect. These are now reported in the paper.
>
> In any case, there is the parameter that changes signs between cities. While number of rooms is significant and positive in Madrid and Barcelona, in Valencia the sign is consistently negative for all models. The negative sign could be due to consumer preferences or other conditions that are specific to Valencia.

3.(d) Finally, I am wondering how the authors estimated the spatial model with matrices of size 44,270 x 44,270 etc. I assume that the authors use a concentrated log-likelihood approach.

> We estimated the spatial models with the `spatialreg::lagsarlm` function (R package {spatialreg}) which relies on maximum likelihood estimation of the spatial simultaneous autoregressive lag model $y = \rho\, Wy + X\, \beta + e$. According to the documentation "ρ is found by `optimize()` first, and β and other parameters by generalized least squares subsequently"; in other words, yes, the approach uses the concentrated log-likelihood. The spatial weight matrices are large but sparse; after revising the matrices to specify the weights using the inverse of the distance (see below), neighbors of order higher than 6 are small and can be filtered. With these conditions, estimation takes only a few minutes in a laptop with the following specs:

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
## $vendor_id
## [1] "GenuineIntel"
##
## $model_name
## [1] "Intel(R) Core(TM) i9-9980HK CPU @ 2.40GHz"
```

```
##
## $no_of_cores
## [1] 16

## 34.1 GB

## [1] "RAM: 34.1 GB"
```

Fourth Comment

4.(a) I like the horse race among the different models including the spatial model. However, in case of the spatial autoregressive model (SAR), the exogenous variables can no longer be interpreted as causal effects. Moreover, it is not clear to me how this spatial specification affects the estimate parameter for the market segments. In particular, there might be a kind of interaction between the spatial weight matrix and the indicator for the market segments.

> Thank you for this perceptive comment. Indeed, the variables in models with spatial autocorrelation cannot be interpreted as marginal effect. James LeSage and others have shown how the lagged variable creates a multiplier effect, and that the impacts on the dependent variable consist of a direct impact of the attribute, but also an indirect impact due to the multiplier. The sum of these two gives the total impact, which depending on the sign of the spatial autocorrelation parameter can be smaller or larger than the direct impact.
>
> Comparison of models Market Segments/Spatial/Spatial with Market Segments indicates that the parameters vary depending on whether spatial heterogeneity, or spatial autocorrelation, or both effects are considered. It is well-known that spatial heterogeneity and association can co-exist (e.g., Bourassa, Cantoni, and Hoesh 2007; Paez, Uchida, and Miyamoto 2001), however, the two effects typically cannot be told apart a priori, and their presence is generally inferred from the empirical results. In our case studies, the improved fit suggests that spatial heterogeneity (i.e., the spatial market segments) are a better interpretation of the data than spatial autocorrelation alone, thus showing one possible interaction between spatial autocorrelation (and the weight matrix) and the indicators for market segments. In particular, once that heterogeneity is considered, the strength of spatial autocorrelation is reduced. The second way that the two effects interact is through the multiplier effect of spatial autocorrelation note above. As seen in Tables 3, 5, and 7 of the revised version of the paper, the total effect of each market segment is greater than the direct effect alone.

4.(b)I would like to see a discussion on how the estimates are influenced in their efficiency and consistency. In addition, the authors should test different spatial weight matrices such as distance- and neighborhood-based matrices. For instance, in case of Barcelona and Madrid the six-nearest neighbors definition might not be appropriate because there are areas with missing neighbors in between city areas, which leads to a meaningless long distance to the next direct neighbor.

> This is a great suggestion. For the revisions, we have changed the specification of the spatial weights matrix from nearest neighbors to inverse distance. This has the effect of reducing the influence of more distant neighbors. Interestingly, the results appear to be robust to the specification of the matrix. To be sure, there are some changes in the estimated values of the parameters, but qualitatively the results are the same and none of our conclusions change.
>
> We would like to thank you again for your thoughtful feedback, which has helped to improve the quality of the paper.

References

Bourassa, S. C., E. Cantoni, and M. Hoesh. 2007. “Spatial Dependence, Housing Submarkets, and House Price Prediction.” Journal Article. *Journal of Real Estate Finance and Economics* 35 (2): 143–60.

Du, Mengnan, Ninghao Liu, and Xia Hu. 2019. "Techniques for Interpretable Machine Learning." Journal Article. *Communications of the ACM* 63 (1): 68–77. https://doi.org/10.1145/3359786.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. 2021. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." Journal Article. *The Econometrics Journal* 24 (1): 134–61. https://doi.org/10.1093/ectj/utaa014.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. "Definitions, Methods, and Applications in Interpretable Machine Learning." Journal Article. *Proceedings of the National Academy of Sciences* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

Paez, A., T. Uchida, and K. Miyamoto. 2001. "Spatial Association and Heterogeneity Issues in Land Price Models." Journal Article. *Urban Studies* 38 (9): 1493–1508.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." Journal Article. *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." Journal Article. *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839.