Hedonic Price Modelling with Geographically Weighted Regression in Medelln, Colombia

Sara Arango, Juan Carlos Duque.

***Abstract*—In this paper we use Geographically Weighted Regression to estimate Real Estate Prices with an hedonic pricing approach, accounting for selected contextual and structural variables for city appraisal records of the Medelln, Colombia in 2014, and claiming that this type of analysis can be equally efficient for emergent cities such as Medelln, regardless of the size of their informal markets.**

***Index Terms*—Hedonic pricing, real estate, Geographically Weighted Regression.**

I. Introduction

ESTIMATING the prices of real estate is useful and necessary for taxation, public policy, infrastructure and equality in cities.

Hedonic pricing. Hedonic pricing is an approach that has been naturally linked to real estate as it welcomes the modelling of assets that are usually not valued by traditional pricing strategies, such as natural resources, accesibility, among others.

GWR is a spatial method developed by Fotheringham et al. [2] to account for the fact that independent variables might have different effects on a dependent variable depending on spatial location. For example, the effect of distance to subway stations in a city in real estate prices may depend on geographical factors that might not be evident from other independent variables.

GWR and hedonic pricing.

Medelln is the second city of Colombia, with more than two and a half million inhabitants and a population of over three million in its metropolitan area. It is a focus of development for the country, and construction plays a major role in its dynamics. **CITA SOBRE MEDELLIN SIENDO IMPORTANTE EN REAL ESTATE**.

Most of the formal development in Medelln is very uniform in its types. The developers do not have a clear idea of what customers want, and thus they keep building over the same formula that has worked this far. *Universidad EAFIT* funded a project with its Research Group in Spatial Economics (RiSE) to assess the real estate prices in Medelln, in order to seek for what it is that costumers value in housing, and to set foundations for a spin-off with this prpose.

GWR was used as a method for the hedonic pricing estimation of real estate in Medelln using structural variables (those particular to the housing unit itself) and contextual variables (those related to neighbour quality) to predict total price. An important amount of geographical processing had to be done in order to collect, process and standardize this information, typically unavailable for the real estate business to use.

Some questions arose:

- How successfully can the price per unit area be modeled? Since Area does explain most of the total price, it would be smart to predict square meter prices, as it could be a more robust measure that can overcome how much area shadows other variables in the estimation of the total price. Predictability would be lost, but capability to understand what the public really want could be gained.
- What contextual variables can be meaningful to improve the actual estimation of real estate prices in Medelln?
- Does the geographical location of a housing unit absorb other important variables in hedonic pricing theory? How does it relate to them? Which can better explain the price?

Sara Arango Center for Urban Science and Progress, *CUSP*. New York University, United States.

Group in Spatial Economics, *RiSE* Universidad EAFIT, Medelln, Colombia

II. Methodology

A. Data and variables

A first important step of this project was to process geographical data from different sources to add to the original data base that was used and analyzed with Universidad EAFIT.

1) Housing data was obtained from OIME, Medellns municipal Real Estate Market Observatory. It can be found online 1 . Initially, it has 2003 observations for new and used houses and apartments across the city of Medelln in the year of 2014 (See Figure 1.).
   These observations come from appraisals for taxation purposes. They present a reasonable spatial distribution. Notice that there are not many data fields across the center of the city (along the middle line crossing from south to north) because that is a restricted zone for construction, and the less dense zone of the city for residential uses.
   The analysed dataset comprises the following variables. The summary data was calculated after removing atypical values and observations with missing data (absolute z-scores greater than 3):
   - Stratum. Categorical index that serves as an indicator for income. Ranges from one to six. In Spanish, from lower to higher income: 1 Uno (90 counts); 2- Dos (253 counts); 3- Tres (479 counts); 4-Cuatro (246 counts); 5- Cinco (179); 6 Seis (114).
   - Parking. Boolean factor reflecting whether a residential unit has at least one parking space or not. Counts: 1120 False and 241 True.

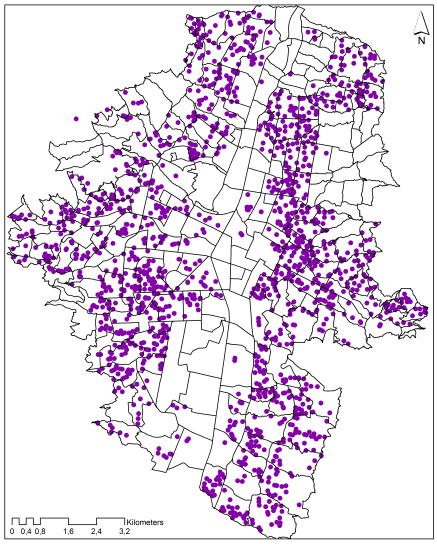

Fig. 1. Distribution of data points.

TABLE I
SUMMARY OF THE NEWLY CREATED VARIABLES.

value	**Density**	**Open space**	**Transit**	**Downtown**	**Roads**
mean	24630	0.001	0.0113	0.040	0.002
std	14576	0.001	0.0077	0.019	0.002
min	0	0.000	0.0006	0.001	0.000
25%	14453	0.000	0.0055	0.027	0.001
50%	21218	0.001	0.0099	0.039	0.001
75%	32645	0.001	0.0153	0.052	0.003
max	77837	0.04	41.2	0.085	0.013

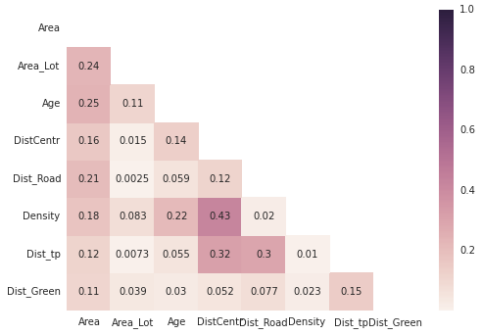

Fig. 2. Distribution of data points.

- Utility room. Boolean variable indicating individual storage rooms. Counts: 944 False and 417 True.
- Price. In COP. By November of 2015, One US Dollar costs around three thousand Colombian pesos.
- Area. In squared meters.
- Age. In years. Medelln is a relatively new city that has undergone exponential growth especially since the 1960s, without much interest in architectural preservation.
- Spatial coordinates. X and Y coordinates of data points in WGCS 1984 projected system.

2) In order to account for contextual variables (those that could indicate added value for relatively similar properties given their distance to natural features, public spaces or transportation networks) data was obtained from:
   - Medellns department of Planning Zoning plan 2 (ArcGIS geodatabase). Includes road networks, metro and BRT stations locations, existing parks, among others. A good amount of geographical processing was needed in order to create variables for nearest distance to features for each of the data points, and to extract the features that were needed in the case of existing green parks and only the main roads in the network.
   - Neighbor density, given by the Life Quality Survey, performed by the city of Medelln. This polygon data was extracted for each of the data points, depending on their geographical location.
     May be this is what you are after:

B. Methods

III. RESULTS

A simple Spearmans correlation shows an interesting story (See Figure 2). As it will be discussed, most of the total price of the sampled houses can be predicted by their total area. It can be seen that Area has a relatively higher than usual degree of correlation to other variables compared to the rest of variable to variable relationships, although the three highest correlations obtained were:

- Density and Distance to Downtown are relatively highly positively correlated. This is only a demonstration of the known fact that most of Medelln lives in its periphery 2.
- Distance to Mass Transit and Distance to Downtown: Mass transit is focused in certain central corridors where most of the people do not actually live. This is an indicator of the lack of accessibility of this system.
- Distance to roads and distance to mass transit. This is mainly explained by the fact that the mass transport system overlays the main roads network.

Figure 3. displays the relationship between price and other critical variables, colored by strata. There seem to be different behaviors for different strata. This will be further explored below.

By comparing the critical variables mentioned below, some regressions were performed, and it was noted that Density creates multicolinearity issues and does not add to the R2 value, although it is found as meaningful in a simple regression that includes Area. Distance to green areas also was found not to be significant in a simple regression and also creates multicolinearity with unit area. The same behavior was observed for distance to roads and distance to public transport.

According to this, the extra geographical features that were added, are not useful for increasing prediction analysis in a global model such as ordinary linear regression. As it will be discussed below, the explanation for this might be that these contextual variables might only be meaningful when local behaviors are taken into account. So far, with just Strata and Area, most of the variance of the price was already described,

and adding more variables only seems to increase the noise and complexity.

The variable lot area was also removed for its correlation to the total area.

Figure

Regression analysis: Using a simple feature selection algorithm that ignores the aforementioned discussion about correlation yields the results displayed in **Table 3**. Only the presence or absence of a Storage room is not significant with a significance level of 95%. Distance to mass transit presents a p-value higher than the rest, but still lower than 0.05 for the stated level of confidence. The independent variables have the expected impacts: Age decreases the value, higher strata imply higher priced properties, parking ads to the value of the unit, distance to mass transit increases the price of the property. This last observation makes sense in a city like Medelln, where the mass transit system is not very efficient and goes through areas that are traditionally congested, loud, polluted, among others.

The same simple feature selection model for Price per Unit Area yields the results displayed in Table 4. It is worth noticing that a Principal Component Analysis was used for both of the models and did not yield any significant increase in prediction capabilities nor any evident cluster more evident than what is already indicated by strata. The relationship between price per squared meter and area is not that evident, this can be shown by the significant yet negative coefficient in the regression in Table 4, and by the plots displayed in Figure 4. This is consistent with the prior belief that modeling the squared meter can be somehow more informative, and, surprisingly, the estimation capability was not lost at large.

The possibilities displayed here need to be more refined in order to create a robust model that can be used for more general cases in Medelln. It is important to notice that, since the city is very segregated itself, Strata can pick many behaviors that otherwise could be attributed to geographical location.

TABLES OF REGRESSION RESULTS

Relationship between variables

Geographically weighted regression: Although the purpose of this exercise was not to develop a robust brand new geographically weighted regression model, the fact that all the contextual variables created for this exercise were not picked up as important for a global regression, an interest to see how a local model would behave under these conditions. ArcGISs GWR method was used only for Area and Green Spaces. Strata is expected to be picked up by the geography itself.

Since inference is complicated in GWR estimation, no assertions can be made, only that the expected added value to green areas according to location has the expected behaviour that is indeed observed in Figure 7: wealthier areas are more attracted to this feature, whereas poorer areas actually repel it because it might usually mean less access to services.

tables

PLOT

IV. Discussion

Not necessarily a novel approach, but it is a novel exercise in the context of emerging cities with potentially very obscure sub markets.

It might be worth just modelling subsets of the city.

GWR must be taken and understood with care since it is a very novel technique and validation is harder, still to be more rigorous.

What is the best way to standardize? what is the best way to treat variables?

Feature selection.

V. Conclusions

- A basic clustering approach does broadly reveal that the submarkets that are expected from the understanding of the city are the ones that are revealed by the prices.
- The effect of the contextual variables used for this analysis is negligible for a global linear model. Their effect might be masked by a Simpsons paradox like effect, and thus the importance of geographically weighted regression or other local method. This result might not mean that these variables should be discarded for Geographically Weighted Regression, only that global analyses are not suitable for increasing prediction of real estate prices to a very high degree.
- Using the square meter price of housing units can be a useful and more robust measure than total price because of the latters sensitivity to total area. This needs to be further explored when a complete GWR model is developed.

VI. Conclusion

The conclusion goes here.

Appendix A
Geographical inputs

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank the City of Medelln's Real Estate Observatory (OIME) in the head of Juan Pablo Barrero.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Fotheringham, A. Stewart., Chris Brundson, and Martin Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, West Sussex, England: Wiley, 2002. Print.