GeoDa Center
for Geospatial Analysis
and Computation
Arizona State University

Is the Price Right? Assessing Estimates of Cadastral Values for Bogota, Colombia

Nancy Lozano-Gracia and Luc Anselin

2011

Working Paper Number 03

Is the Price Right? Assessing Estimates of Cadastral Values for Bogotá, Colombia

Nancy Lozano-Gracia[*] Luc Anselin[†]

Abstract

Hedonic house price models are increasingly applied in the process of mass appraisal, in which econometric specifications are used to obtain automated valuation of properties for taxation purposes. The predictive quality of such models is important, since it directly affects the revenue stream of local authorities. In this paper, we assess the relative predictive performance of different model specifications used in automated valuation. Specifically, we focus on the issue of spatial heterogeneity by comparing models that utilize different definitions of housing submarkets. In addition, we consider the inclusion of "spatial" explanatory variables in the form of distance to various amenities as computed from a GIS. We apply this to data from the city of Bogotá, Colombia, a pioneer in the application of mass appraisal techniques in a developing country context. We find that specifications that include the submarkets improve predictive performance and that the inclusion of the spatial variables is superior to the traditional models of homogenous zones. However, even the best models are still characterized by relatively poor performance in the form of a high degree of overprediction of the house value. In addition, the predictive performance of the models varied by socio-economic stratum in the city, which suggests that the dynamics of the housing markets in these strata would require closer and separate attention. These results may provide further guidance to enhance mass appraisal practice in the city of Bogotá as well as potentially other Latin American cities.

[*] Finance, Economics and Urban Unit, Sustainable Development Network, World Bank, Washington, DC. nlozano@worldbank.org

[†] GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ. lanselin@asu.edu

1 Introduction

Taxes on land and property are an important source of revenue for local authorities. Their level, design, and control are key elements of local fiscal policy. In this context, it is important to keep in mind that property tax revenues are determined not only by the tax rate, but also by the estimated value of the property. To operationalize the levy of property taxes therefore requires the periodic (typically, annual) valuation of a large number of properties. For large cities, manual appraisal becomes impractical and instead automated procedures for so-called mass appraisal are increasingly used. In mass appraisal, an econometric model specification forms the basis for obtaining a prediction of the value of a property. Careful design of the model specification and estimation methods are important for maintaining the credibility and political acceptability of the cadastral office. The objective is to implement automated procedures that reflect the actual property value as closely as possible.

Traditional approaches for mass valuation of land and real estate parcels can be classified into three main categories: market value comparison, income capitalization, and cost approach. The market value comparison approach analyzes data from sales of similar parcels to determine the value of units being appraised. Although this is generally the preferred approach, it is strongly dependent on the number of market transactions available and the accuracy of the sales prices. Lack of accurate data is one of the main obstacles to the practical implementation of this method.

A second approach measures the present value of future benefits obtained from owning the parcel. Most commonly the value is assessed using actual or estimated income derived from a property in combination with the application of a capitalization factor. This approach is particularly useful when market data on comparable parcels sold is not available.

A third technique is the cost approach. This is based on estimation of the replacement costs of a property after subtracting depreciation of improvements. The market value of land (as if vacant) is added to the improvements to obtain the total assessed market value of a property. The quality of this approach depends on the availability of information on construction costs, depreciation rates, and land values. Construction costs may be calculated from other known building costs, from published statistics or expert estimates.

The consensus is that the market value approach is preferred because it captures the value of the construction and land as well as that of amenities. In many instances, the valuation of amenities reflects government expenditures and policies. In other words, the market methodology takes into account, at least to some extent, the value of the location and local context. In the cost approach, these factors are embedded in the land value.

Once a valuation methodology is selected, regression analysis can be used to estimate the taxable value for each property. In this context, hedonic house price models are increasingly applied for both mass appraisal as well as mortgage underwriting in developed countries, such as the US, the UK, Canada, the Netherlands, Spain and Switzerland, among others. More recently, Latin American countries, and Colombia in particular, have also started using hedonic models for mass appraisal purposes. However, the question remains whether conclusions derived for cities in the developed world apply to cities in developing countries. To date, there is very little empirical evidence to that effect.

In this paper, we provide some initial results in this respect and consider the case of the city of Bogotá, Colombia. Our objective is to improve the specification of the hedonic model by adding a spatial perspective. We consider two different aspects. The most important one is taking into account spatial heterogeneity, in the form of housing submarkets. We include different definitions and assess how they affect the predictive performance of the models. We also consider "spatial" explanatory variables in the form of distance measures to various amenities that can be readily calculated from a GIS and thus can be automated. This contrasts to the collection of actual property characteristics, which is still very labor intensive. Using extensive data for residential properties in the city of Bogotá, Colombia, we compare the different model specifications in terms of their out-of-sample predictive performance.

The remainder of the paper is organized in four sections. We start by providing a general introduction to hedonic price models and discuss the role of space and the implications it may have on the hedonic price equation and its estimation. In Section 3 we provide a general description of the data. Section 4 summarizes the main results. Finally, we formulate some concluding remarks in Section 5.

2 Background

In the mid 1960s, Lancaster (1966) provided the micro-economic basis for a new branch of utility theory in which utility was defined not as a function of goods, but rather as a function of the characteristics of the goods. His work primarily focused on the consumer, and it is not until Rosen (1974) that a model of market equilibrium was formulated that would take into account both consumers and producers. Rosen (1974) modeled what could be considered different goods as essentially one commodity, allowing the utility of an individual to be a function of the characteristics of the commodity, and producer costs to be dependent on the type of the good. In the context of a housing market and house sales transactions, the hedonic price equation defines the market equilibrium, after all interactions between supply and demand have taken place. The basic model (e.g., Freeman 1999) assumes that the utility of a household or an individual is a function of a composite good x; a vector of location-specific environmental characteristics q; a vector of structural characteristics S; a vector of social and neighborhood characteristics N; and a vector of locational characteristics L. The hedonic price function is then an equilibrium price equation where the price of house i is defined as a function of the house characteristics:

$$P_i = P(q_i, S_i, N_i, L_i), \tag{1}$$

The hedonic equilibrium equation holds for all properties within the same market. In practice, this creates a problem when different submarkets may be present in the same urban area. This is a special case of so-called spatial heterogeneity, where significant differences in the parameters between submarkets could lead to misleading inference, when ignored (see, e.g., Anselin 1990, Anselin and Lozano-Gracia 2009). Consequently, accounting for differences between submarkets should improve both model interpretation as well as the predictive ability of the model. In a mass appraisal context, the latter is the main goal, hence it is important to address the potential existence of submarkets.

In this regard, there are two concerns. On the one hand, submarkets may be identified to ensure that properties are close substitutes, and, as a result, that the prices for house characteristics equalize in equilibrium. This implies that a different price equation should be estimated for each submarket. However, when the properties in a submarket are too homogeneous, the use of the associated estimates for out-of-sample prediction may not be optimal (Bourassa et al. 2003). In other words, in the context of mass appraisal for a cadastral update process, it is important to select the submarkets in such a way that the accuracy of overall prediction is improved, rather than focusing on the specific differences in coefficients across submarkets.

The treatment of the delineation of submarkets in a housing market is well developed in the literature. Suggested methods include the use of political boundaries (Goodman 1981, Goetzmann and Spiegel 1997, Brasington and Hite 2005), the definition of areas by appraisers (Bourassa et al. 2005), and the application of statistical techniques, such as principal components and cluster analysis (Bourassa et al. 2003, 1999), model based clustering (Day et al. 2004), and hierarchical models (Goodman and Thibodeau 1998, 2003). As expressed in Bourassa et al. (2003), the consensus of previous studies is that geographically defined submarkets improve appraisal performance. However, these conclusions are limited to studies that pertain to developed economies, and little is know about how this would apply in a developing context.

The city of Bogotá, Colombia, is a pioneer in Latin America in terms of updating its cadastral valuation systems. Specifically, as part of the 2008 and 2009 cadastral updates, the Bogotá Cadastre employed hedonic models for its mass appraisal that included both locational and neighborhood characteristics. The introduction of explicit spatial variables in the mass appraisal process exploited a recently updated GIS information data base for the city. This allowed the construction of new variables that expressed the distance to selected sites of interest, thereby quantifying accessibility to amenities.

Historically, before the availability of GIS data for the city, neighborhood characteristics were collected through a very labor-intensive and costly process. Individuals hired by the cadaster physically walked through the city, and classified each neighborhood using a combination of observation and their own intimate knowledge of the city. Criteria used in the classification included categories for the main activities, access to public services, and dominant land use. Neighborhoods were thus classified into *homogeneous* zones (HZ). The collection of information required for defining the HZ zones in Bogotá currently represents about 73 % of the total costs of estimating cadastral values. This mostly consists of payments for temporary employees, assessors, and assessing firms involved in the process (Ruiz and Vallejo 2010). Since the total cost of the updating process in 2009 was around 7.8 million U.S. dollars, the cost related to the definition of HZ zones in Bogotá alone can be estimated to be close to 6 million U.S. dollars. Clearly, any gains in efficiency in the model specification that would avoid this costly process are to be given serious consideration.

The categorization into homogenous zones is traditionally used for two purposes: first, to estimate land values and second, to capture neighborhood heterogeneity by including fixed effects variables in the hedonic specification used for estimating construction value.

After successfully updating the cadastral value of all properties in the city, the Cadastre office is trying to modernize and reduce the cost of maintaining the cadastral data base up to date. It wants to introduce new methodologies that streamline the process and reduce the need for massive fieldwork and labor intensive-operations. Specifically, the Cadastre wants to move toward the use of market values in the process of estimating property values as a basis for property tax, as is the case in OECD countries (Ruiz and Vallejo 2010). However, as in many developing countries, transaction data is limited, and when available, it usually suffers from under-reporting due to tax avoidance. The current methodology circumvents this problem by estimating cadastral values through individual assessments of a statistical sample of properties. It is important to understand how the models that were used in the updating process of 2008 and 2009 perform when applied to the new *market* data.

In our study, we are particularly interested in two aspects of the hedonic specification. First, we consider the relative performance of models that account for separate housing submarkets. Specifically, we consider four alternative specifications of submarkets. Second, we assess the comparative performance of the traditional model specification with homogenous zones relative to the inclusion of explicit spatial (distance) variables obtained from a GIS. This comparison is important, since the GIS-based variables are easy to compute automatically, whereas the traditional approach is highly labor-intensive and prone to subjective assessment. Given the high degree of public controversy associated with mass appraisal, there is great interest in developing a system in which subjectivity is removed as much as possible.

We estimate each model using data for residential properties in Bogotá and base our assessment on a comparison of out-of-sample predictive performance.

3 Data and Methods

The data used in this study were collected and provided by the Unidad de Analisis Economico del Catastro Distrital (UAECD) unit of the Bogotá Cadastre.[1] House characteristics and house values are based on Cadastral records. House values are the values on the books used for tax purposes. They are the result of appraisal efforts carried out over different time periods. Consequently, since it is not practical to update all the house value data annually, historical data must be properly inflated to result in comparable prices. Specifically, we transformed all values into 2008 pesos. In addition, we included a control variable in the hedonic specification that represents the year of the property's last update. Property values are estimated as the sum of building value plus land value. Econometric models may be employed to estimate the building values, but by law land values can only be determined by appraisers based on the concept of homogeneous zones.[2]

In Bogotá, as in most developing cities, comprehensive market data on property transactions is not available. To alleviate this problem, in 2002 the city Cadastre office started an initiative towards the collection of *market* data, under a program called the *observatory*. Under the auspices of the *observatory* program, data from property sales are collected through a city wide census-like process in which all sales announced through signs or local newspapers are recorded. Cadastre officials survey these properties and contact the owners, and pretending to be potential buyers, negotiate on the house price, to try to get as close as possible to what would actually be a sales price. Negotiations are based on cash payments. Consequently, the market values obtained through this process are not from actual sales, but rather they are adjusted asking prices.

Information on asking prices is combined with Cadastral records that have information about property characteristics. The resulting data set includes information on both characteristics and asking prices of 14,079 properties that were for sale between 2002 and 2007, with all values transformed to 2008 Colombian pesos. In addition to data on structural characteristics of the house, spatial variables are included in the form of the distance from each property to important landmarks such as parks, malls and wetlands. All characteristics included in the estimation of our house price equations are listed in Table 1, with descriptive statistics given in Table 2.

[1] This roughly translates to Unit of Economic Analysis of the District's Cadastre.

[2] The estimation of land values is governed by Resolution 2555 of 1988 which mandates that land prices should be determined through the survey process described above for the definition of homogeneous zones. The value of building and improvements of a property may be estimated by means of econometric models, but the land value may not, due to this legal constraint. This is an interesting issue, especially in the context of potential modifications of current practice. Any recommendation that fundamentally alters this practice would require a revision of the law. For the purposes of the current analysis, we limit the hedonic analysis to the estimation of full value of the properties.

Table 1: **Description of Variables**

Variable	Description
Dependent Variable	
lnvalue	Log of value per squared meter in Cadastre books (pesos)
General House Characteristics	
area	Total area of property in squared meters
Age	Years since construction completion
Points	Points given by appraisers as general measure of quality (0-100)
Indicator Variables	
armazon es111	Structure Pre-built material
armazon es112	Structure Brick
armazon es114	Structure Concrete up to 3 levels
armazon es115	Structure Concrete up to 4+ levels
adec alineada	Slope “flat”
adec empinada	Slope “steep”
vias peato	Pedestrian routes
estado viasbueno	Good roads
propiedad horz	Horizontal Property
cubierta est131	Roof: made of waste or asphalt
cubierta est133	Roof : provisional cover
cubierta est134	Roof: simple cover, ceramic tiles
cubierta est135	Roof: aluminum and tiles
cubierta est136	Roof: good cover, water proof coat
conserv est142	Structure in acceptable shape
conserv est143	Structure in good shape
conserv est144	Structure in excellent shape
fachada211	Facade Poor
fachada213	Facade Acceptable
fachada214	Facade Good
muros est000	No wals
muros est121	Walls built from tiles or Adobe
cubierta221	Walls with no cover
cubierta222	Walls with pressed brick cover
pisoacab231	Floor: dirt
pisoacab232	Floor: cement or wood
pisoacab235	Floor: acrylic, granite, tiles, rubber
pisoacab236	Floor: Parquet, carpet or marble
pisoacab237	Floor: high end marble or other luxuries
tno coci411	No kitchen
conserv coci441	Kitchen in bad shape
enchape coci4212	Unfinished kitchen or covered with cement or common tile
enchape coci423	Kitchen walls covered with small tile or paper
enchape coci424	Kitchen walls covered with large tile

Table 1. Continued

Variable	Description
tno bano314	Large bathroom
enchape bano321	Unfinished bathroom
conser bano341	Bathroom in bad shape
mobil bano000	No furniture in bathroom
mobil bano331	Basic furniture in bathroom
conser bano3434	Good and luxury furniture in bathroom
Neighborhood Characteristics	
Urban 7576	Indicator for urban designation: park
Urban 5152	Indicator for urban designation: improvements
area act72a75	Indicator for protected area (natural resources)
Actd predt3	Residential area (general)
Actd predt2	Residential area (specialized)
Autocons	If unit built by owner
Minimum Distances (km)	
d1f djuana	Waste Disposal Site (Dona Juana)
d1f estr1	Strata 1
d1f ccmed	Medium size mall
d1f humed	Wetlands
d1f milit	Military Base or Station
d1f moteles	Hourly Motels
d1f parqmetro	Metropolitan park
d1f parqzon	Zonal Park
d1f aereop	Airport

Table 2: **Descriptive Statistics**

Variable	Mean	S.D.	Min	Max
Indicator Variables				
pisoacab235	0.2996	0.4581	0	1
pisoacab236	0.0922	0.2893	0	1
pisoacab237	0.0056	0.0744	0	1
tno coci411	0.0251	0.1566	0	1
conserv coci441	0.3209	0.4668	0	1
enchape coci4212	0.3338	0.4716	0	1
enchape coci423	0.2280	0.4195	0	1
enchape coci424	0.2220	0.4156	0	1
tno bano314	0.0137	0.1162	0	1
enchape bano321	0.0590	0.2356	0	1
conser bano341	0.3165	0.465	0	1
mobil bano000	0.0213	0.1443	0	1

Table 2. Continued

Variable	Mean	S.D.	Min	Max
mobil bano331	0.0826	0.2754	0	1
conser bano3434	0.1420	0.3490	0	1
Neighborhood Characteristics – Homogenous Zones Data				
Urban 7576	7.98e-06	0.0028	0	1
Urban 5152	0.0955	0.2939	0	1
Area act72a75	0.0024	0.0485	0	1
Actd predt3	0.0024	0.0494	0	1
Actd predt2	0.0796	0.2706	0	1
Autocons	0.6458	0.4783	0	1
Distances				
d1f djuana	14695.54	7266.142	115.83	33248.4
d1f estr1	1971.75	1361.01	0	6824.25
d1f ccmed	2158.29	1599.66	0.0008	16035.35
d1f humed	4246.23	3424.11	1.5434	20871.52
d1f milit	2983.52	1529.23	8.2267	12446.55
d1f moteles	1342.66	838.24	0.000016	9500.81
d1f parqme o	2852.59	1668.50	6.4446	12518.14
d1f parqzon	1110.74	665.77	4.7774	7467.27
d1f aereop	10421.11	4269.62	906.44	28338.64

The point of departure in our analysis is what we refer to as the base model. In this specification, the estimated coefficients are the same for all properties, irrespective of their location. We introduce spatial heterogeneity into this model in two different ways. In the first, we introduce fixed effects (FE) to represent submarket heterogeneity. In the second, we estimate a separate model in each submarket.

We consider four different specifications of submarkets. Two of these are based on the so-called strata determined by the National Planning Department. This results in a classification of residential areas into six groups, based on similarity in access to services, utilities, quality of roads, etc. The rating ranges from 1 for the lowest stratum to 6 for the highest. All housing units in the neighborhood are allocated to the same stratum, resulting in a degree of spatial clustering. Also, in Bogotá, neighborhoods tend to be strongly segregated by income, which results in broader spatial areas representing a stratum. We consider two ways to incorporate this into the definition of submarkets. In one, the six strata are combined into two overall groups, one consisting of the lower strata (1–3), the other of the upper strata (4–6). In the other, we consider six submarkets, one for each stratum. Their spatial distribution is illustrated in Panel (a) of Figure 1.

A third definition of submarkets is based on the administrative units of the city. This definition pertains to the localities that were created with the intention of decentralizing the provision of services and administrative functions. Currently, there are twenty such administrative units in the city. Among these, the locality of Sumapaz is mainly comprised of natural reserves and thus does not contain any residential units. It is therefore excluded from the analysis. As a result, there are a total of nineteen submarkets based on localities. Their spatial layout is illustrated in Panel (b) of Figure 1.

A fourth definition of submarkets is based on a categorization of the type of real estate, referred to as activities. This designation stems from the survey efforts associated with the definition of the *homogeneous zones* and yields three categories: specialized residential, general residential, and non-residential. This classification is inherently non-spatial as it pertains only to the particular characteristics of the property. Therefore, this does not yield spatially delineated submarkets, but rather individual properties classified into one of the three categories. The spatial distribution of these properties is illustrated in Panel (c) of Figure 1. While there is some degree of clustering, this definition of submarket should be considered to be non-spatial. Table 3 provides a summary of the five specifications considered.

Table 3: **Definition of Submarkets**

Name	Definition
Base	All observations pooled into one market
2-SMK	Two submarkets defined using strata: strata 1–3 and strata 4–6
6-SMK	Six submarkets, corresponding to strata 1 through 6.
19-SMK	19 submarkets corresponding to 19 administrative units (localities)
3-ACT	Three submarkets defined by the main activity in the area

We evaluate the four submarket definitions using fixed effects in a common model specification. We contrast this with models where the coefficients are allowed to vary between submarkets, i.e., a so-called spatial regime specification. Finally, we consider the difference between the homogeneous zone specification and the inclusion of spatial (distance) variables in the model.

All models are estimated using ordinary least squares (OLS) applied to a subset of a random selection of 90% of the 14,079 observations. The resulting coefficients are then used to predict the value for the 10% out of sample observations. For each model specification, this is repeated for 100 random samples in order to remove the influence of a particular sample selection. The predictive performance of the models is based on a comparison of the average performance over the 100 out-of-sample predictions.

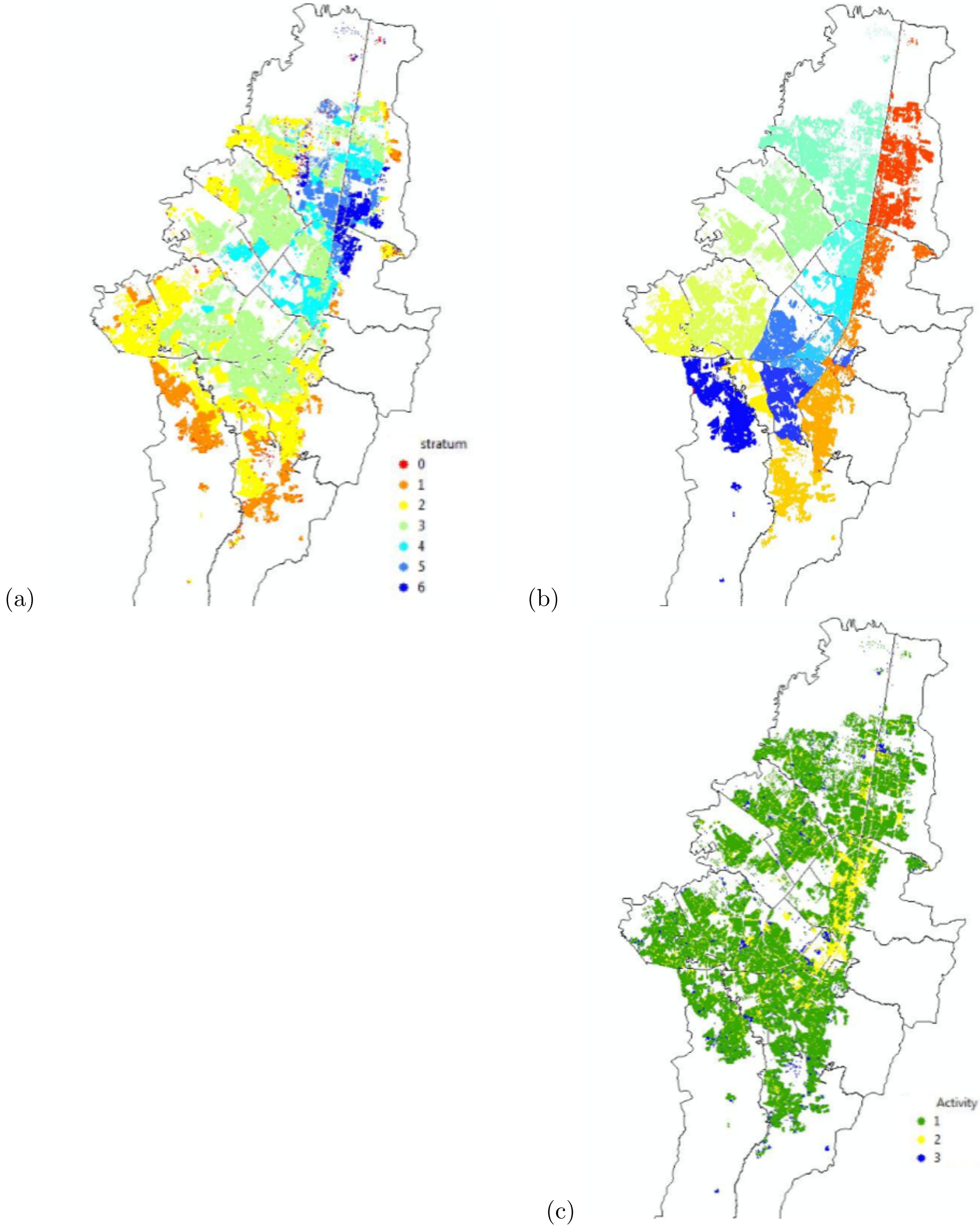

Figure 1: Spatial Distribution of Alternative Submarket Definitions. Panel (a) Strata, Panel (b) Localities, Panel (c) 3 Activities

4 Results

We focus on the relative predictive performance of the different model specifications based on the accuracy of the out-of-sample predictions for each model. Since each model is estimated 100 times, each time for a different subsample, the out-of-sample set used for prediction is also different each of the 100 times. We summarize the individual model fit by a median measure of accuracy across the 100 predictions. Specifically, we will consider the median average absolute percentage error, and the median percentage of predictions that fell within respectively 10% and 20% of the true value. In addition, we include the median percentage overpredictions, since politically, overprediction is considered more harmful than underprediction. Given our attention to predictive performance, we are less interested in the magnitude and significance of the individual regression coefficients. Also, given the large number of estimates, it would be impractical to list them all. To illustrate the type of results we obtained, we include a sample estimation result in Table 7 in the Appendix. In the remainder of the Section, we devote our attention to comparative predictive accuracy.

There are two dimensions to this comparison. First, we consider different specifications for a base model, in which the differentiating characteristics of submarkets are accounted for as fixed effects (FE) through indicator variables. In this context, we are particularly interested in a comparison between specifications that include the traditional, survey based neighborhood characteristics used in the designation of homogeneous zones (HZ), and those where distance to various facilities (computed by means of a GIS) are included. The second dimension of comparison considers so-called spatial regimes, where a separate set of estimates is obtained for each of the submarkets. These estimates are then used to obtain predicted values for the out-of-sample units, using the value of an indicator variable to allocate the units to their proper submarket model.

The comparison of the models is not straightforward. Not only is there considerable variation in the coefficient estimates and in-sample model fit among the 100 subsamples, but there is additional variation introduced in the prediction subsample. However, by carrying out this number of experiments, we avoid conclusions based on an arbitrarily selected subset and provide a more realistic assessment of what we can expect in an actual automated appraisal exercise.

The base model and its fixed effects variants are compared in Table 4. The median (across 100 subsamples) average absolute percentage error ranges from high of 36.81% for the Base model without any FE, to a low of 32.8% for the most complete specification with fixed effects for the six strata as well as the 19 localities and with the distance variables included. The latter is a marginal improvement over the model with just the strata and localities fixed effects (32.96%). In sum, the inclusion of the full specification results in a 4% improvement in prediction relative to the Base model. However, it should also be noted that this performance is far from impressive. Interestingly, while the inclusion of the HZ variables improves prediction for the Base model and the specifications with respectively Strata FE and Localities FE, it does not for the model with

Table 4: **Median Predictive Accuracy: Base Model with FE**

Model	Average Absolute Percentage Error	% Overprediction	Predictions within 10%	Predictions within 20%
No HZ variables				
Base	36.81	73.63	17.06	32.94
Strata FE	33.88	77.90	18.05	34.61
Localities FE	35.06	76.12	17.24	33.40
S and L FE	32.96	78.75	18.02	34.33
S, L FE and Distance	32.80	79.25	18.02	33.83
Including HZ variables				
Base	35.61	75.87	17.09	32.94
Strata FE	33.49	78.71	18.19	34.33
Localities FE	34.50	76.76	17.48	33.30
S and L FE	33.17	79.25	17.95	34.26

both included. Without a better quantification of the precision of these assessments it is not possible to suggest "significance", but it is useful to note that the model with fixed effects and distance slightly outperforms its counterpart with HZ variables.

The ranking of models on percent overprediction runs counter to that on median average absolute percentage error. The most fully specified models rank worst (79.25% overprediction for both the model with all FE and respectively distance and HZ), whereas the simplest Base model rates best (73.63%). Also, there seems to be a systematic bias towards overprediction. This is worrisome for policy purposes, since overprediction is worse than underprediction. Apparently, all models tend to overvalue the unit characteristics within-sample relative to the out-of-sample data.

Yet another gauge of the predictive performance is provided by the percent predictions within respectively 10 and 20% of the true value. Here, the model with just Strata FE seems to perform best. Including HZ with Strata FE yields 18.19% for the 10% range and 34.33% for the 20% range, while the Strata FE model without HZ obtains 18.05% for the 10% range and 34.61% for the 20% range. The fully loaded model with distance is comparable to its HZ counterpart, slightly better for the 10% range and slightly worse for the 20% range. All FE models do better than the Base model. In other words, while the Base model overpredicts slightly less, its on-target performance for smaller prediction errors is not as good as the FE models. The distance model is marginally better than the HZ model in this respect.

In Table 5 we compare the specifications that explicitly allow for spatial regimes in the form of submarkets with different coefficient estimates for each. To facilitate the comparison with the fixed effects models, we repeat the results for the "best" such model as the top row in the Table. This specification includes fixed effects for Strata and Localities as well as the distance variables. In terms of median average absolute percentage error, the performance is very similar, and slightly better for all submarket specifications, with 29.99% for 19-SMK as best. All the

Table 5: **Median Predictive Accuracy: Submarkets**

Model	Average Absolute Percentage Error	% Overprediction	Predictions within 10%	Predictions within 20%
	Base Model with FE			
S, L FE and Distance	32.80	79.25	18.02	33.83
	Submarkets Models			
2-SMK	32.46	69.58	22.67	42.54
6-SMK	31.42	60.59	26.40	49.32
19-SMK	29.99	67.91	24.41	45.27
3-ACT	32.38	79.21	17.91	33.62

submarket models are within 2.5% of each other. A much greater improvement is seen in the percent overprediction, which is reduced by almost 20% for the 6-SMK model relative to the Base FE specification (60.59% compared to 79.25%). Only the non-spatial 3-ACT submarket model does not seem to be an improvement in this respect (79.21%). In terms of prediction with the 10 and 20% range, the 6-SMK model is again best, yielding an improvement of more than 8% relative to the Base FE model in the 10% range (26.4% compared to 18.02%) and of more than 15% in the 20% range (49.32% relative to 33.83%). The 19-SMK model only performs slightly worse in this respect. Interestingly, the three "spatial" submarket models always do better on these three prediction criteria than the non-spatial one (based on the traditional housing unit classification).

Table 6: **Percentage of Overprediction by Model and Stratum**

Variables	Strata					
	1	2	3	4	5	6
	Base Model with FE - no HZ variables					
Base	90.16	79.85	72.54	70.63	65.09	45.32
Strata FE	69.89	76.05	80.21	80.21	76.24	66.67
Localities FE	83.90	79.80	77.35	74.43	61.02	50.35
S and L FE	71.36	78.06	80.62	80.92	76.44	67.13
S, L FE and Distance	71.13	79.09	81.26	82.16	76.36	67.87
	Base Model with FE - including HZ variables					
Base	84.89	79.92	77.52	72.04	65.69	43.64
Strata FE	69.69	78.32	80.95	80.74	76.40	66.90
Localities FE	80.50	79.28	79.30	75.26	62.33	50.00
	Submarkets					
2-SMK	64.02	70.60	73.19	66.22	64.63	57.39
6-SMK	58.43	58.87	61.46	60.81	62.98	53.33
19-SMK	61.98	66.53	67.06	71.10	72.50	71.43
3-ACT	72.22	79.61	80.91	81.77	75.00	67.21

As a final comparison, we consider the percentage overprediction for all the models classified by the stratum the out-of-sample unit belongs to. Again, the results are the median percentage across the 100 subsamples. They are listed in Table 6. The results illustrate the complexity associated with achieving reasonable out-of-sample predictive performance in the presence of heterogeneity across submarkets. For example, the Base model without any fixed effects scores between 90.16% overprediction in the lowest stratum and 45% overprediction in the highest stratum. While the former is the worst across all models and all strata, the latter is second best (only the Base model with HZ variables does better at 43.64% also for Stratum 6). In other words, the traditional base model (with and without HZ) does well for the highest stratum, but does very poorly for the others, especially for the lower strata. On the other hand, the three SMK models, and especially 6-SMK, do much better across the range of lower strata. Of the 12 specifications considered, the 6-SMK model has the lowest percentage overprediction for strata 1 through 4, is third for stratum 5 (slightly behind two localities FE models) and is fourth for stratum 6.

The variability of the predictive performance of the different specifications across strata raises an important question about the overall approach. Clearly, neither the FE nor the regimes models fully capture the heterogeneity present in the submarkets and further model refinement seems called for. While the models that explicitly account for regimes, and especially 6-SMK seem to do better than the other across a range of strata, the performance remains uneven, which warrants further investigation. It also highlights that a policy based on "one size fits all" can results in serious side effects that impact social equity.

5 Concluding Remarks

In this paper, we set out to assess the relative predictive performance of a range of econometric specifications for hedonic house price models to be employed in an automated mass appraisal system. We focused in particular on the context of a large metropolitan area in a developing country, exploiting a rich database of over 14,000 residential properties from the Cadastre office in Bogotá, Colombia. We were interested in the effect of an explicit consideration of submarket heterogeneity in the models as well as on specifications that included traditional *homogenous zones* (HZ) versus those that employed distance variables computed from several GIS layers.

Even though our conclusions are limited by the nature of the sample used and the experiments carried out, we nevertheless gained some useful insights. First, it was clear that models that include distance variables perform as well and in some instances even slightly better than the traditional specification with HZ. This is important from a policy perspective, since its automated implementation could result in significant cost savings. Second, models that allowed for heterogeneity of coefficient estimates across submarkets outperformed those that implemented heterogeneity by means of fixed effects. The specification that included six spatial submarkets based on socio-economic strata performed best. In contrast, the inclusion of heterogeneity in the non-spatial 3-ACT model did not improve upon the fixed effects specifications. Third, we observed considerable variability in the degree of overprediction of all models across strata. This is potentially worrisome, since this has clear equity implications. Again, the 6-SMK model did best for the lower strata and slightly worse than the best for the highest stratum. This is an aspect of heterogeneity that requires further investigation.

What do our findings mean in terms of whether the “price is right”? Overall, the median out-of-sample predictive performance was less than stellar for all models. We found substantial support for the inclusion of distance variables rather than the traditional HZ. We also found substantial support for the expression of submarkets as regimes rather than as fixed effects. However, the degree of overprediction of the models remains a concern. If this indeed reflects a “best practice” performance, then significant discounting of the predicted values obtained in an actual automated mass appraisal exercise would seem to be in order.

The models included in this exercise did not consider spatial autocorrelation explicitly. While it would have been near impossible to implement this across all models and all subsamples, an improved predictive performance may be achievable by using geostatistical techniques such as kriging. This will be investigated in future work.

Acknowledgments

The authors wish to acknowledge financial support from the Lincoln Institute of Land Policy and the GeoDa Center for Geospatial Analysis and Computation. We also want to thank Bogota’s Cadastre office for providing the data. Special thanks are due to Gabriel Vallejo for his constant support with the data and for providing feedback at different stages of the work. Previous versions of this paper were presented at the Lincoln Institute of Land Policy Latin American Annual Conference (San Jose, Costa Rica, 2009) and a seminar series at the Inter-American Development Bank (October, 2010). We thank the participants for their comments. The usual disclaimer applies.

References

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30:185–207.

Anselin, L. and Lozano-Gracia, N. (2009). Spatial hedonic models. In Mills, T. and Patterson, K., editors, *Palgrave Handbook of Econometrics: Volume 2, Applied Econometrics*, pages 1213–1250. Palgrave Macmillan, Basingstoke, United Kingdom.

Bourassa, S., Hamelink, F., Hoesli, M., and MacGregor, B. (1999). Defining residential submarkets. *Journal of Housing Economics*, 8:160–183.

Bourassa, S., Hoesli, M., and Peng, V. (2003). Do housing submarkets really matter. *Journal of Real Estate Finanace and Economics*, 35:143–160.

Bourassa, S. C., Cantoni, E., and Hoesli, M. (2005). Spatial dependence, housing submarkets, and house prices. Research Paper 151, International Center for Financial Asset Management and Engineering.

Brasington, D. M. and Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Regional Science and Urban Economics*, 35:57–82.

Day, B., Bateman, I., and Lake, I. (2004). Nonlinearity in hedonic price equations: an estimation strategy using model-based clustering. Working paper, Center for Social and Economic Research of the Global Environment, University of East Anglia, UK.

Freeman, A. M. I. (1999). *The Measurement of Environmental and Resource Values*. Resources For the Future, Washington, DC.

Goetzmann, W. N. and Spiegel, M. (1997). A spatial model of housing returns and neighborhood substitutability. *Journal of Real Estate Finance and Economics*, 14:11–31.

Goodman, A. C. (1981). Housing submarket within urban areas: definitions and evidence. *Journal of Regional Science*, 21(2):175–185.

Goodman, A. C. and Thibodeau, T. G. (1998). Housing market segmentation. *Journal of Housing Economics*, 7:121–143.

Goodman, A. C. and Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12:181–201.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74:132–156.

Rosen, S. M. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82:534–557.

Ruiz and Vallejo, G. (2010). Using land registration as a tool to generate municipal revenue: Lessons from Bogotá. Annual Bank Conference on Land Policy and Administration, World Bank, 1818 H St NW, Washington, DC.

6 Appendix

Table 7: **Sample Hedonic Regression Results**

Variables	Base Model	Strata FE	Localities FE	S and L FE	S, L, and Distances
Constant	16.2***	1.6.3***	1.8***	16.63***	17.1***
Area	0.0029***	0.0026***	0.0028***	0.0026***	0.0026***
Age	0.0069***	0.0021***	0.0056***	0.0022***	0.0016***
Points	0.0437***	0.0224***	0.0329***	0.0195***	0.0187***
armazon es111	0.0593	0.0200	-0.0519	-0.0931	-0.102*
armazon es112	-0.0250	-0.0356	-0.0507*	-0.0567*	-0.0602
armazon es114	-0.162***	-0.123***	-0.114**	-0.0972**	-0.0741
armazon es115	-0.178***	-0.0816***	-0.150***	-0.0833***	-0.0893E***
propiedad horz	0.0396**	-0.0463***	0.0472***	-0.0272*	-0.0291**
cubierta est131	-0.122	-0.118	-0.0280	-0.0561	-0.0677
cubierta est133	-0.0543***	-0.0040	-0.00374***	0.00032	-0.0008
cubierta est134	0.0500**	0.0126	0.0834***	0.0263	0.0276
cubierta est135	-0.0092	-0.0747**	0.0588*	-0.0515*	-0.0543*
cubierta est136	-0.0407	-0.0207	0.0594	0.01195	0.0202
conserv est142	-0.126 ***	-0.0689***	-0.0103***	-0.0709***	-0.0665***
conserv est143	-0.0996***	-0.0453***	-0.1.12***	-0.0593***	-0.0549***
conserv est144	-0.0036	0.0249	0.0181	0.0320	0.0503
fachada211	0.120***	0.0597**	0.0800***	0.0445*	0.0387
fachada213	0.0178	-0.0224	0.0336**	-0.0116	-0.0059
fachada214	0.0882***	-0.0270	0.142***	0.0012	0.00237
muros est000	0.956***	0.241	0.826***	0.2557	0.2660
muros est121	-0.0929	-0.125	-0.0014	-0.0355	0.0181
cubierta221	-0.0397	-0.0442	-0.0378	-0.0481*	-0.0426
cubierta222	-0.0025	0.0182	0.0099	0.0175	0.0196*
pisoacab231	-0.165**	-0.139**	-0.136**	-0.1447**	-0.0127**
pisoacab232	-0.104***	-0.0291*	-0.0556***	-0.0284*	-0.0212
pisoacab235	-0.0290**	-0.0230*	-0.0264**	-0.0164	-0.0171
pisoacab236	-0.0491**	-0.0430**	-0.0260	-0.0328	-0.0316
pisoacab237	-0.102	-0.0855	-0.0653	-0.0663	-0.0696
tno coci411	0.0263	-0.0612	-0.0412	-0.0764**	-0.0901**
conserv coci441	0.0744***	0.06***	0.0679***	0.0528***	0.0542***
enchape coci4212	-0.0397	0.0013	-0.0490	0.0059	0.0027
enchape coci423	-0.0601	0.0022	-0.0761**	0.0005	-0.0055
enchape coci424	-0.151***	-0.0290	-0.1.27***	-0.0200	-0.0241
istno bano314	0.0320	0.0396	0.0381	0.0406	0.0530**
enchape bano321	-0.0642**	-0.0389	-0.0478*	-0.0447	-0.0394
conser bano341	-0.0249	-0.0314*	-0.0077	-0.023241	-0.0197

***Significant at 1%, ** Significant at 5%, * Significant at 10%

Table 7. Continued

Variables	Base Model	Strata FE	Localities FE	S and L FE	S, L, and Distances
mobil bano000	-0.826***	-1.12 ***	-1.03 ***	-1.1905***	-1.24***
mobil bano331	0.0941***	0.0460**	0.0678***	0.0385**	0.0316*
conser bano3434	-0.230***	-0.162***	-0.168***	-0.1437***	-0.145***
autocons	0.0344***	0.0599***	0.00009	0.0404***	0.03***
uso38	-0.201***	-0.226***	-0.2.24***	-0.2218***	-0.220***
d1f djuana					-0.0007
d1f estr1					0.0087*
d1f ccmed					-0.0193 ***
d1f humed					-0.0245***
d1f milit					-0.0422***
d1f moteles					-0.0159**
d1f parqmetro					0.0089
d1f parqzon					-0.0362***
d1f aereop					-0.0014
R-squared	0.685	0.741	0.72	0.752	0.7573
Adjusted R-squared	0.685	0.7405	0.7194	0.7514	0.7559

***Significant at 1%, ** Significant at 5%, * Significant at 10%