(2022) 52, 321–343

The Review of Regional Studies

The Official Journal of the Southern Regional Science Association

A House Price Modeling Based on Clustering and Kriging: The Medellín Case *

Hernán D. Villada-Medina[a], Juan F. Rendón-García[a], César A. Ramírez-Dolores[b], and Gerardo Alcalá[c]

[a]*Facultad de Ciencias Económicas y Administrativas, Instituto Tecnológico Metropolitano, Colombia*

[b]*Instituto de Energías Renovables, Universidad Nacional Autónoma de México, México*

[c]*Centro de Investigación en Recursos Energéticos y Sustentables, Universidad Veracruzana, México*

Abstract: In this study, house prices are modeled using a mixed two-stage model for mass appraisal employing valuations of second-hand housing units conducted in Medellín, Colombia. In the first stage, submarkets of houses that share non-spatial attributes are created using clustering; in the second stage, the spatial dependency is incorporated into the house price estimation using kriging. The best results were obtained when the sample was divided into three submarkets using property area and age as the classification criterion and later applying a Matérn kriging model to submarket 1, a spherical kriging model to submarket 2, and a circular kriging model to submarket 3. These results may provide further guidance to enhance mass appraisal practice in other Latin American cities as well as potentially other cities in developing countries.

Keywords: clustering; geostatistics; k-means; kriging; mass appraisal; value map.

JEL Codes: C21, R12, R32

1. INTRODUCTION

House prices are estimated individually or collectively. According to Ling and Archer (2018), traditional approaches for individual appraisal of real estate can be classified into three main categories: the cost approach, the income approach, and the sales comparison approach. The cost approach is used mainly to appraise newly built properties, and the income approach is for commercial and investment real estate. The sales comparison approach consists of comparing the prices of a sample of residential properties sold in the area of the subject

*We wish to thank two anonymous referees for providing helpful comments. All remaining errors are our own. Villada-Medina and Rendón-García are Full-time Professors at Instituto Tecnológico Metropolitano, Colombia. Ramírez-Dolores is an Assistant Professor at Universidad Nacional Autónoma de México, México, and Alcalá is a Full-time Professor at Universidad Veracruzana, México. *Corresponding Author:* Gerardo Alcalá, E-mail: galcala@uv.mx

ISSN 1553-0892, 0048-49X (online)
www.srsa.org/rrs

property with similar non-spatial properties (area, bathrooms, age, bedrooms and so on). These two conditions of similarity (non-spatial and spatial) are not always easy to meet and represent the main obstacle for the appraiser. However, this approach has the advantage of being easily understood by buyers and sellers thanks to its simplicity. Moreover, the sales comparison approach is regarded as the most reliable approach because it captures not only the value the construction and the land but also the value of public and private amenities, which in many cases reflects the government expenditures, policies and the social status of the neighborhood (Appraisal Institute, 2013; Ling and Archer, 2018).

Valuing several properties at the same time (i.e., mass appraisal) requires the development of statistical models to automate such a process, while individual appraisal demands the direct intervention of an appraiser who visits the property to determine its value. Mass appraisal seeks to determine the market value of properties in order to define tax policies, set prices in simultaneous purchase and sale transactions, or draw up property value maps of cities or regions for urban development decision-making. For this reason, there is a current interest in models that make it possible to value groups of properties in a quick, accurate, and affordable manner.

For purposes of mass appraisal, hedonic models using the classical method of multiple linear regression (MLR) have been widely implemented. These models try to estimate the value of a property (dependent variable) according to two or more attributes (independent variables) such as area, age, number of rooms, and number of bathrooms. In addition, to capture the spatial effect on housing prices, these models also include the spatial distance to places of interest (e.g., shopping malls, educational institutions and tourist attractions).

This paper follows the proposal of Can (1992), who states that house prices depend on two variables: non-spatial attributes (e.g., lot size, type of construction and age) and neighborhood attributes (e.g., land use and externalities). However, instead of analyzing both variables at the same time (as in traditional hedonic models), house prices in Medellín, Colombia, are modeled by combining a data mining method (clustering) and a geostatistical technique (kriging). The hypothesis of this study is that, within the context of mass appraisal, house prices can be modeled by separately evaluating non-spatial attributes (through an appropriate selection of submarkets and clustering criteria) and spatial attributes (through an appropriate selection of kriging models). By applying this methodology, the benefits of the sales comparison approach for individual appraisal, previously described, will be obtained for mass appraisal too. On the one hand, the clustering technique allows that the value of the construction of a property is only compared with those properties with similar non-spatial attributes (unlike traditional hedonic models, which use all the samples at the same time). On the other hand, the kriging regression (based on the location) captures not only the value of the land but also the value of public and private amenities.

This paper suffers the limitation of using a sample of just 0.03% (293 residential properties) of the housing occupied units in the city. The sample size is one of the main drawbacks of conducting this type of research in countries such as Colombia. Mainly due to the current security conditions in the country, the prices of properties negotiated in the Colombian real estate market are not registered in the Public Instruments Registry Offices or the Property Registry Offices. However, the article is still valuable for the application of alternative techniques (different from traditional hedonic models) to the real estate market of a city in a

developing country like Colombia, since most related studies using computer-assisted mass appraisal (CAMA) have focused on cities in developed nations (Lozano-Gracia and Anselin, 2012; Wang and Li, 2019). This study also sheds light on the effect of implementing submarkets (in terms of their number and clustering criteria) on the accuracy of price prediction in mass appraisal. In addition, it examines the behavior of different kriging models applied to the spatial interpolation of property values.

This paper is divided into five sections, including the introduction: Section 2 is a literature review, Section 3 describes the methodology used in this study, Section 4 presents and discusses the results, and Section 5 draws some conclusions.

2. LITERATURE REVIEW

The most widely used mathematical model applied to mass appraisal is hedonic pricing. In this model, all the attributes that affect the value of a property are jointly analyzed, usually through multiple regression (Lai, 2011) and the Ordinary Least Squares (OLS) method (Cebula, 2009; Monson, 2009; Teixeira et al., 2010). Nevertheless, using the traditional econometric approach to examine spatial data is problematic due to the spatial correlation that occurs when there are levels of spatial dependence between the variables. This is particularly true for the real estate market, in which properties with high and low prices tend to be concentrated in specific areas (Anselin, 1988; Basu and Thibodeau, 1998). Hedonic models often try to solve this problem of spatial dependence by including the distance from the property to the city's downtown in the group of explanatory variables using Von Thünen's theory (Stevens, 1968) and/or dummy variables associated with a certain classification of the area or neighborhood where it is located (Cellmer et al., 2014; Montero et al., 2018).

The development of artificial intelligence techniques and Geographic Information Systems (GISs) has produced a series of studies that use methods such as Artificial Neural Networks (ANNs) (McCluskey et al., 2012; Selim, 2009; Mimis et al., 2013; Vo et al., 2015), decision tree models (McCluskey et al., 2014; Reyes-Bueno et al., 2018), and clustering (Gabrielli et al., 2017; Napoli et al., 2017). Moreover, several studies on the application of machine learning to mass appraisal have been published and obtained successful results using the random forest method (Antipov and Pokryshevskaya, 2012; Čeh et al., 2018; Credit, 2021).

Despite the success of some machine learning models in mass appraisal in residential real estate (Wang and Li, 2019), most of them do not account for spatiality in the data. Therefore, some efforts can be found in the literature that try to solve this problem by mixing or complementing machine learning technics with traditional models or existing models. Examples of these combinations are: ANN and GIS (García et al., 2008); quantile regression forest (QRF) and kriging Córdoba et al. (2021), and also multicriteria analysis and genetic algorithm (Morano et al., 2018).

The works of Calka and Bielecka (2016) and Calka (2019) combine clustering and kriging for mass appraisal purposes (as it is done in the current article). These two studies divide the market into clusters (local submarkets) based only on the property's non-spatial (structural) attributes and then perform interpolation for each cluster separately using ordinary kriging. The method is applied in both cases to the city of Siedlce (Poland), and results show that

the estimation error for a property's value, using the mean absolute percentage error, does not exceed 8.5% and 10%, respectively. For these works, spherical variograms are used, yet in the present work, the Gaussian, spherical, Matérn, and spherical models are used, along with a sweeping process for the lag width and cutoff distance parameters of the empirical variogram, in order to obtain the best-fit kriging model.

However, although these studies have generally reported better results compared to traditional hedonic models, it is still not clear whether these benefits actually justify the use of more complex techniques. For instance, Bourassa et al. (2007) point out that the gains, in terms of accuracy, of including dummy variables for submarkets in an OLS model are superior to any other advantage of using geostatistical or lattice methods such as Spatial Autoregressive (SAR) models.

Finally, it is worth noting that all the studies mentioned so far have been done in Europe, the United States, Canada, and China. This shows the lack of empirical evidence about whether or not the conclusions derived for cities in the developed world apply to cities in developing countries (Lozano-Gracia and Anselin, 2012).

3. METHODOLOGY

3.1. Two-Stage Price Modeling

In this study, house prices in Medellín (Colombia) are modeled by combining a data mining method (clustering) and a geostatistical technique (kriging), which is a procedure that differs from traditional hedonic models that analyze both types of variables simultaneously. In the first stage, submarkets (in which residential properties share non-spatial attributes) are created through clustering. In the second stage, spatial dependence is incorporated into house price estimation by means of kriging.

3.2. Data Collection

Medellín is the second-largest industrial city in Colombia. With an area of 380.64 km^2, it is 1.5 km above sea level and located in the region known as Valle de Aburrá. It is crossed, from South to North, by the Medellín River, which determines the configuration of its natural landscape due to its geoforms, topography, and hydrological features. In addition, numerous streams that flow towards such rivers divide the valley sides where its urban areas continue to grow.

A census by the National Administrative Department of Statistics (abbreviated DANE in Spanish) in 2018 reported that Medellín had an estimated population of 2.4 million inhabitants. The city is divided into 16 *comunas* (districts), 5 *corregimientos* (townships), and 271 neighborhoods. Additionally, according to the DANE (2019), there are 892,151 residential properties in the city: 57.63% of them are apartments; 39.82%, houses; 2.47%, rooms; 0.02%, ethnic dwellings; and 0.06%, other types of dwelling.

This study used data from 293 residential properties that were valued between 2014 and 2019. Figure 1 shows two maps: the location of Medellín in Colombia and the distribution of the residential properties used as the sample in this study. Each appraisal included

information concerning the date of the appraisal, property price, number of rooms, number of bathrooms, built-up area, age of the property, and geographic location. Table 1 presents the descriptive statistics of the dataset. It should be noted that, in this case, as in all cases of mass appraisal, attribute selection depends on the available information sources.

Figure 1: Location of Medellín in Colombia (Left) and Distribution of the Residential Properties under Study (Right).

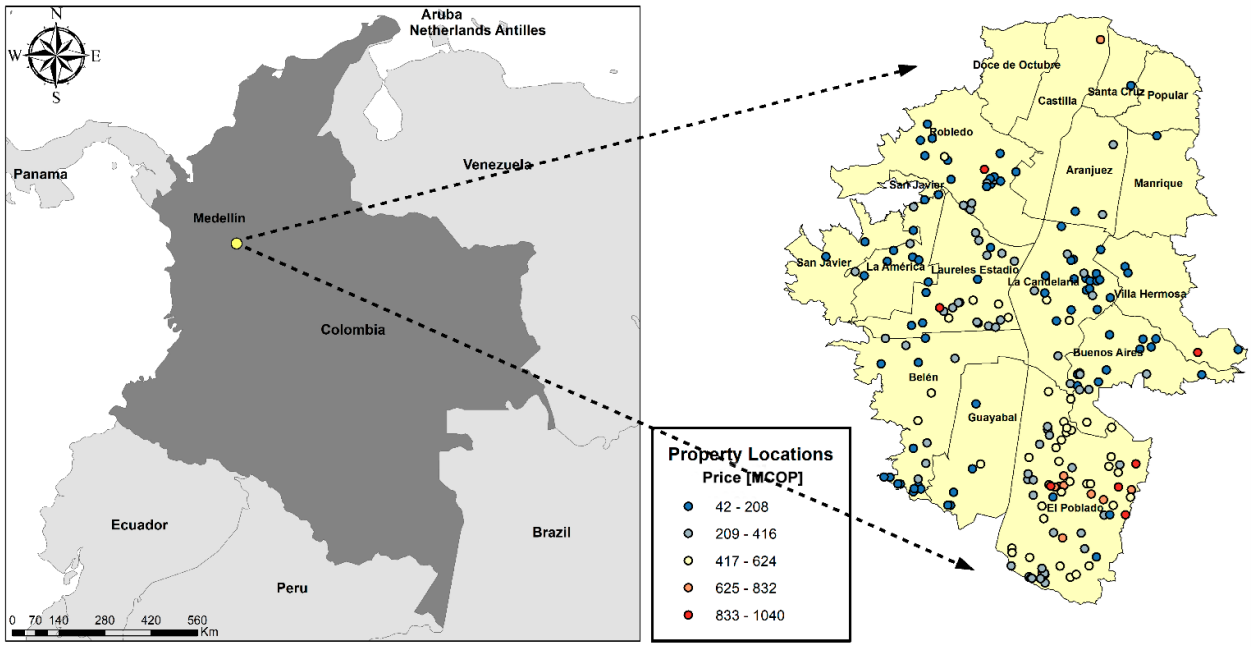


Table 1: Descriptive statistics of the data set

Variable	Min	Max	Mean	Std. Dev
Price [COP*]	42,000,000	1,040,000,000	275,800,000	216,439,656
Area [M^2]	30.00	278.00	95.74	53.33
Number of Rooms	1	5	2.74	0.76
Number of Bathrooms	1	5	1.97	0.76
Age [Years]	0	53	15.56	11.94
Type	Apartment	House		
Number of Observations	249	44		

*Colombian Pesos.

These data were provided by Lonja de Propiedad Raíz de Medellín y Antioquia, a local real estate association, from its private database, because the prices of properties traded in this market are not recorded in the Registry Offices of Public Instruments or the Land Registry Offices due to the country's current security conditions. For this reason, the values set forth in the appraisals are a good approximation to the real market prices of the properties.

In addition, they are better than the offer prices constantly used to conduct this type of research in Colombia, since the latter are typically overpriced to allow room for negotiation in the event that a deal can be struck[1].

3.3. Clustering

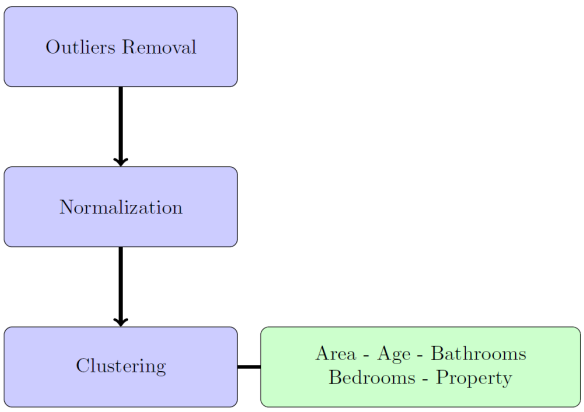
As mentioned earlier, in the first stage of this study, a location-insensitive model was developed based solely on the non-spatial attributes of the properties. Using a data mining technique (i.e., k-means clustering), clusters of similar properties that created submarkets were formed. Each cluster was characterized by defined attribute values expressed as items in a rating scale. In this model, the independent variables are built-up area, age of the property, number of bathrooms, number of rooms, and type of property (apartment or house).

The k-means algorithm has been widely used to divide n points found in a d-dimensional space into k groups (Vattani, 2011). Given a set of observations $Z_1, Z_2, \ldots, Z_n$, where each observation is a real d-dimensional vector, the k-means algorithm aims to partition n observations into $k \leq n$ clusters, $S = \{S_1, S_2, \ldots, S_k\}$ to minimize the within-cluster sum of squares. Its purpose is to solve:

$$\arg\min(S) \sum_{i=1}^{k} \sum_{Z \in S_i} \|Z - \mu_i\|^2 \tag{1}$$

where μ_i is the mean of the points in S_i. When individual properties are grouped, the spatial dimension (d) is defined by the number of attributes of the properties under analysis. Since the k-means algorithm is sensitive to the number of clusters adopted *a priori*, hierarchical agglomerative clustering was applied using Euclidean distance combined with Ward's method (Ward, 1963).

Figure 2: Clustering Process



[1]Property taxes in Medellín are defined by the cadastral value of the properties which is defined by the Treasury of Medellín. The local real estate association that provides us the data is a completely independent organization, and their appraisals have nothing to do with the owners' taxes.

3.4. Kriging

The kriging method estimates regionalized variables $Z(\boldsymbol{x}_0)$ at unsampled locations $\boldsymbol{x}_0$. In addition, it can be expressed as a linear combination of all available measurements of Z according to its general equation (de Marsily, 1984; Marko et al., 2014)

$$\hat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^{n} \lambda_0^i Z(\boldsymbol{x}_i), \tag{2}$$

where $\hat{Z}(\boldsymbol{x}_0)$ represents a kriged value at location $\boldsymbol{x}_0$; and λ_0^i, the values of the weighting factors assigned to individual observations $Z(\boldsymbol{x}_i)$ at location $\boldsymbol{x}_i$.

In order to produce optimal estimates, the estimated variables must be unbiased and have minimum variance. These two conditions make it possible to obtain weighting factors λ_0^i that lead to the following linear system with $(n+1)$ unknowns (de Marsily, 1984):

$$\sum_{j=1}^{n} \lambda_0^i = 1, \tag{3a}$$

$$\sum_{j=1}^{n} \lambda_0^i \gamma(\boldsymbol{x}_i - \boldsymbol{x}_0) + \mu = \gamma(\boldsymbol{x}_i - \boldsymbol{x}_0); \qquad i = 1, 2, ..., n \tag{3b}$$

where μ is a Lagrange multiplier; and $\gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)$, the semi-variogram, which depends on the spatial separation distance ($\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$) between two points ($\boldsymbol{x}_i$ and $\boldsymbol{x}_j$), this distance is referred to as lag. This indicates that the values of the weighting factors, $\lambda_0^i = 1$, only depend on the separation distances between the individual observation points provided by the semi-variogram.

Therefore, an estimate of the variogram is required for the geostatistical estimation (kriging). It is usually obtained by computing $\gamma(\boldsymbol{h})$ for discrete lags and then fitting a suitable lag function to these estimates. The most widely used estimator of the variogram is Matheron's estimator. It states that the value of the empirical variogram for a separation distance of $\boldsymbol{h}$ is half the average squared difference between a target value $Z(\boldsymbol{x}_i)$ at some sample location and the value $Z(\boldsymbol{x}_i + \boldsymbol{h})$ of the neighbor at a distance $\boldsymbol{x}_i + \boldsymbol{h}$ (Mehrjardi et al., 2008):

$$\gamma(\boldsymbol{h}) = \frac{1}{2n(\boldsymbol{h})} \sum_{i=1}^{n(\boldsymbol{h})} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h})]^2, \tag{4}$$

where $n(\boldsymbol{h})$ is the number of data pairs within a given class of distance and direction. The presence of a spatial structure where observations $Z(\boldsymbol{x}_i)$ and $Z(\boldsymbol{x}_i + \boldsymbol{h})$ (close to each other) are spatially autocorrelated will result in $\gamma(\boldsymbol{h})$ values that are small compared to those of far apart uncorrelated pairs of points (Lark, 2000; Mehrjardi et al., 2008). It should be noted that this estimator is asymptotically unbiased for any intrinsic random function. However, since it is based on squared differences among data, it is very sensitive to outlying values of Z. In fact, a single outlier can distort the estimate of the variogram because it occurs in several paired comparisons over many or all the lag intervals (Lark, 2000). For this reason,

outliers are removed, and data are normalized during clustering (see Figure 2) before the experimental variogram is computed (Mehrjardi et al., 2008).

The calculations of the empirical variograms will depend on the selected lag distance (D_L), cutoff distance (D_C), and prevailing direction. The cutoff distance is the maximum distance up to which point pairs are considered, while the prevailing direction is associated with the spatial structure of a particular variable at different directions of the field. In this study, the directional aspect of the spatial dependence was not taken into account, as this requires a large number of samples, which leads to isotropic variograms (i.e., no major direction) (Nayanaka et al., 2011).

In this paper, five theoretical models—exponential, Gaussian, spherical, circular (Johnston et al., 2001), and Matérn (Minasny and McBratney, 2005) denoted by Equations (5a), (5b), (5c), (5d) and (5e) respectively—are proposed to fit the empirical variograms. Also, their parameters (nugget, range, and sill) are determined to characterize the spatial dependencies (structures) of different house prices.

$$\gamma(\boldsymbol{h}) = \theta_S \left[1 - e^{-\frac{3\|\boldsymbol{h}\|}{\theta_r}}\right] \tag{5a}$$

$$\gamma(\boldsymbol{h}) = \theta_S \left[1 - e^{-3\left(\frac{\|\boldsymbol{h}\|}{\theta_r}\right)^2}\right] \tag{5b}$$

$$\gamma(\boldsymbol{h}) = \begin{cases} \theta_S \left[\frac{3}{2}\frac{\|\boldsymbol{h}\|}{\theta_r} - \frac{1}{2}\left(\frac{\|\boldsymbol{h}\|}{\theta_r}\right)^3\right] & \text{for } 0 \le \|\boldsymbol{h}\| \le \theta_r \\ \theta_S & \text{for } \theta_r < \|\boldsymbol{h}\| \end{cases} \tag{5c}$$

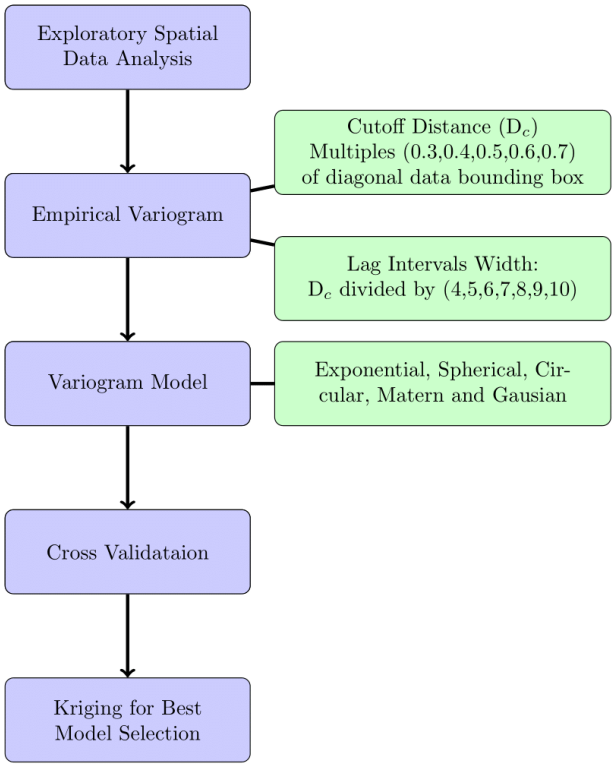
$$\gamma(\boldsymbol{h}) = \begin{cases} \frac{2\theta_S}{\pi}\left[\frac{\|\boldsymbol{h}\|}{\theta_r}\sqrt{1 - \left(\frac{\|\boldsymbol{h}\|}{\theta_r}\right)^2} + \arcsin\frac{\|\boldsymbol{h}\|}{\theta_r}\right] & \text{for } 0 \le \|\boldsymbol{h}\| \le \theta_r \\ \theta_S & \text{for } \theta_r < \|\boldsymbol{h}\| \end{cases} \tag{5d}$$

$$\gamma(\boldsymbol{h}) = \theta_S \left(1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\|\boldsymbol{h}\|}{\theta_r}\right)^{\nu} K_\nu\left(\frac{\|\boldsymbol{h}\|}{\theta_r}\right)\right) \tag{5e}$$

where $\theta_S \ge 0$ is the partial sill parameter; and $\theta_r \ge 0$ is the range or distance parameter that measures how fast the correlations decay with distance. In the Matérn model, ν represents the smoothness parameter; K_ν, a modified Bessel function of the second kind ν; and Γ, the gamma function. Kriging was performed for each cluster, as described in Figure 3. This process can be summarized in five steps: (1) Exploratory Spatial Data Analysis (ESDA), (2) empirical variogram calculation, (3) variogram model fitting, (4) model validation, and (5) generation of maps for the best model (Johnston et al., 2001; Mehrjardi et al., 2008).

The ESDA in Step 1 includes basic statistics for each cluster (submarket), such as the mean and standard deviation of prices, the number of paired comparisons among data, their separation distances, or data density (see Table 7). In Step 2, multiple empirical variograms are constructed for a given cluster by varying the values of the cutoff distance (D_C) and lags. The cutoff distance values are considered multiple values of the diagonal bounding box distance. The lag interval widths, are fractions of the cutoff distance (see Fig. 3). Moreover,

Figure 3: Geostatistical Analysis Process for a Given Cluster



each computed empirical model is fitted according to all the different variogram models (Step 3). Finally, in Step 4, the best map is obtained via cross-validation.

3.5. Estimation Accuracy

The quality of a map is best assessed by comparing the estimated $\hat{Z}$ values with actual observations Z at validation points using an independent (control) data set. When no control data set is available, prediction models are usually validated via cross-validation. In other words, the original point set is divided into two data sets (calibration and validation), and then the analysis is repeated (Hengl, 2009).

In this study, we employed Leave-One-Out Cross-Validation (LOOCV), where each sampling point is used as validation data. The best model for experimental variogram fitting is selected based on R^2 criteria. However, other relevant validation indices are also computed, such as the Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) (Bishop and Lark, 2008; Hengl, 2009; Marko et al., 2014; Calka, 2019). In fact, the best R^2 consideration leads to minimal RMSE, and reduces MAPE values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)\right)^2}{\sum_{i=1}^{n} \left(Z(\boldsymbol{x}_i) - \bar{Z}\right)^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - \hat{Z}\left(\boldsymbol{x}_i\right)\right)^2} \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Z\left(\boldsymbol{x}_i\right) - \hat{Z}\left(\boldsymbol{x}_i\right)}{Z\left(\boldsymbol{x}_i\right)}\right| \tag{8}$$

4. RESULTS AND DISCUSSION

Kriging interpolation was applied to three cases (Case A, Case B, and Case C), and data observations were grouped in sets of two, three, and four submarkets (clusters), respectively. In all the cases, five variables (area, age, number of bedrooms, number of bathrooms, and type of property) were used as clustering criteria: the first criterion (one-dimensional) considered the first variable (area); the second criterion (two-dimensional) considered the first two variables (area and age); and so on until the fifth criterion (fifth-dimensional), the last one, took into account all five variables. Therefore, since there are five different clustering criteria for cases A, B, and C, a total of 45 submarkets must be modeled using a variogram function.

Table 2 presents the descriptive statistics of the 45 submarkets. Such statistics include the case (number of clusters), clustering criterion, submarket for each corresponding criterion, a unique ID for each submarket, mean value of the clustering variables, and number of apartments and houses in the submarket. According to this table, the average price using area and age (clustering criteria 1 and 2) is similar in the case of two and three clusters. However, there is a significant change in the average price when the number of rooms is included in the clustering criteria (clustering criteria 3, 4, and 5). This difference is not observed in the case of four clusters.

Table 2 shows the effect of including new variables in the case of two, three, and four clusters. Additionally, all 45 submarkets were modeled using five cutoff distances, seven lag interval widths, and five variogram models (see Figure 3), which resulted in 175 possibilities to model each submarket. For each submarket, LOOCV was employed to select the best variogram model based on R-Squared criteria, as shown in Table 3. Besides the Case, Clustering Criterion, and Submarket fields, Table 3 includes the variogram model, lag width, cutoff distance, the R-Squared, RMSE, MAE, MAPE values, and the Number of Observation Points. The submarket best modeled for each case is highlighted in bold, and its corresponding clustering Criterion rows use grey background.

The results of Table 3 are summarized in Table 4 for the individual cases, as well as for all the cases combined. The latter table includes the percentage of times each model appeared and the average lag width and cutoff distance. In all the cases, the most suitable model for the different clusters was circular (appearing 55.6% of the time), followed by Gaussian, spherical, and Matérn models (15.6%, 11.1%, and 11.1%, respectively). The exponential option was the least common (6.7%). The mean lag width and cutoff distance were 1018 m and 6418 m, respectively.

In Case A (two submarkets), Table 4 shows that the only models that appeared were the

Table 2: Descriptive statistics of the different clusters.

Case	Clustering criterion	Submarket	Submarket ID	Average price [COP]	Average area [m^2]	Average age [years]	Average number of bedrooms	Average number of bathrooms	Type of property apartment-house
Case A 2 submarkets	1	1	1	519,400,000	164.50	21.77	3.19	2.73	74-17
		2	2	166,086,677	64.77	12.76	2.54	1.63	175-27
	2	1	3	519,400,000	140.00	26.43	3.12	2.46	92-27
		2	4	166,086,677	65.44	8.12	2.48	1.63	157-17
	3	1	5	394,000,000	138.13	24.92	3.23	2.46	98-31
		2	6	195,000,000	62.40	8.19	2.35	1.59	151-13
	4	1	7	391,000,000	136.60	23.41	3.19	2.51	110-29
		2	8	185,262,248	58.82	8.47	2.33	1.48	139-15
	5	1	9	400,800,000	135.90	23.61	3.20	2.49	108-31
		2	10	162,980,576	58.99	8.19	2.31	1.50	141-12
Case B 3 submarkets	1	1	11	394,800,000	202.40	21.20	3.29	3.00	27–8
		2	12	167,000,000	131.90	21.37	3.09	2.41	63-12
		3	13	656,500,000	60.51	12.09	2.49	1.59	159-24
	2	1	14	405,446,667	182.80	19.77	3.24	2.86	43-15
		2	15	149,893,490	96.54	30.86	2.90	2.03	59-12
		3	16	593,900,000	64.59	7.45	2.49	1.63	147-17
	3	1	17	207,000,000	174.96	17.89	3.26	2.82	49-16
		2	18	193,146,395	100.16	30.18	3.14	2.08	55-17
		3	19	579,400,000	60.69	7.83	2.33	1.56	145-11
	4	1	20	214,277,778	162.23	25.30	3.20	2.77	71-16
		2	21	177,743,646	53.35	8.39	1.74	1.35	78-5
		3	22	493,300,000	77.31	13.50	3.09	1.82	100-23
	5	1	23	158,168,081	148.00	26.17	3.25	2.57	73-31
		2	24	201,390,715	48.45	9.90	2.03	1.01	66-13
		3	25	421,100,000	80.25	9.58	2.76	2.09	110-0
Case C 4 submarkets	1	1	26	109,749,367	211.30	21.37	3.30	3.00	19–8
		2	27	257,780,079	51.43	10.22	2.29	1.43	115-14
		3	28	679,300,000	88.35	17.23	2.97	1.99	60-13
		4	29	118,561,633	144.80	21.94	3.14	2.61	55-9
	2	1	30	250,069,288	168.40	15.14	3.16	2.72	45-12
		2	31	451,976,562	74.76	24.43	2.82	1.74	42-19
		3	32	587,900,000	60.17	6.24	2.41	1.59	130-8
		4	33	154,327,869	151.10	36.32	3.19	2.60	32-5
	3	1	34	177,529,049	185.00	20.18	3.31	2.95	40-15
		2	35	362,000,000	75.26	8.55	3.09	1.83	79-13
		3	36	597,400,000	53.83	7.92	1.73	1.38	78-4
		4	37	222,174,543	102.19	31.44	3.03	2.09	52-12
	4	1	38	163,084,765	185.00	20.34	3.28	3.06	40-13
		2	39	221,046,875	47.99	9.71	2.00	1.01	66-11
		3	40	607,100,000	78.25	6.94	2.70	2.07	90-2
		4	41	110,457,143	103.60	29.49	3.18	2.07	53-18
	5	1	42	258,226,182	147.48	26.24	3.10	2.65	83-0
		2	43	230,704,225	46.04	8.82	1.57	1.02	50-1
		3	44	442,638,554	73.11	8.92	2.78	1.91	116-0
		4	45	121,425,490	115.90	20.81	3.33	1.93	0-43

circular and Matérn ones (80% and 10%, respectively). In addition, according to Table 3, the best variogram was obtained for clustering criterion 2. For both submarkets, the best fit model was the circular one. Nevertheless, submarket 2 was adjusted to an R-Squared value of 0.5832; and submarket 1, to 0.12. The cutoff distance and lag width of submarket 2 were 7741 m and 968 m, respectively.

In Case B (three submarkets), Table 4 indicates that the circular model was the most common option (40%), followed by the Gaussian, exponential, Matérn, and spherical models (26.7%, 13.3%, 13.3%, and 6.7%, respectively). Additionally, as shown in Table 3, the variogram best described a submarket was obtained for clustering criterion 2. Submarket 3 was best fitted using the circular model (with an R-Squared value of 0.5811), while submarkets 1 and 2 were best fitted using the Matérn and spherical models (with an R-Squared value of -0.1258 and 0.2201, respectively). This negative value suggests that the null model's prediction is more accurate than that of kriging.

Table 3: Models Selected via Leave-One-Out Cross-Validation.

Case	Clustering Criterion	Submarket	Submarket ID	Model	Lag Width	Cutoff Distance	R-Squared	RMSE	MAPE [%]	Observation Points (N_p)
	1	1	1	Matérn	701	5608	0.0037	200821692	33.23	86
		2	2	Circular	645	3871	0.4478	80970818	36.94	177
	2	1	3	Circular	561	5608	0.1200	225096888	53.41	109
Case A		**2**	**4**	**Circular**	**968**	**7741**	**0.5832**	**101771961**	**36.02**	**154**
	3	1	5	Circular	623	5608	0.1816	215563505	49.76	119
2 submarkets		2	6	Circular	774	3871	0.5121	103455828	37.44	144
	4	1	7	Circular	1051	8412	0.1628	210399681	47.76	131
		2	8	Circular	1003	9031	0.4336	95228844	38.16	132
	5	1	9	Circular	1051	8412	0.1813	210224706	47.94	130
		2	10	Matérn	1290	6452	0.4748	93068789	37.46	133
	1	1	11	Gaussian	1202	8412	-0.1893	209728785	28.05	32
		2	12	Gaussian	565	4522	0.2326	127167193	29.02	72
		3	13	Matérn	1290	9033	0.3944	73197888	34.47	159
	2	1	14	Matérn	1930	9650	-0.1258	208438795	26.82	54
		2	15	Spherical	1594	7972	0.2201	86942896	42.34	65
		3	**16**	**Circular**	**1106**	**7743**	**0.5811**	**98650579**	**32.87**	**144**
Case B	3	1	17	Exponential	1206	9650	-0.1006	198375039	26.48	60
		2	18	Gaussian	759	6834	0.2918	86607028	34.97	66
3 submarkets		3	19	Gaussian	1807	9033	0.5287	95304322	39.36	137
	4	1	20	Circular	2103	8412	0.1359	205963996	35.01	82
		2	21	Circular	954	7630	0.5008	93897000	40.65	71
		3	22	Exponential	748	5984	0.4838	100414468	33.73	110
	5	1	23	Circular	623	5608	0.1711	222583902	48.66	97
		2	24	Circular	1908	7630	0.4721	65323203	32.32	66
		3	25	Circular	486	3400	0.5527	105530392	30.75	100
	1	1	26	Spherical	1103	5515	-0.0403	178616456	22.51	26
		2	27	Exponential	430	3871	0.2737	55315971	30.44	110
		3	28	Circular	971	6795	0.3968	91120137	32.20	67
		4	29	Circular	1574	7872	0.0995	159473948	29.25	60
	2	1	30	Spherical	526	3158	0.0605	188054416	26.46	56
		2	31	Exponential	499	3492	0.1456	72819134	41.23	55
		3	32	Circular	717	6452	0.5286	92037204	33.71	119
		4	33	Circular	1402	5608	0.0166	147003301	32.82	33
Case C	3	1	34	Gaussian	981	9814	-0.1942	220260213	29.10	51
		2	**35**	**Gaussian**	**532**	**4787**	**0.6244**	**99754816**	**30.22**	**83**
4 submarkets		3	36	Circular	2225	8902	0.5565	90656313	38.97	71
		4	37	Spherical	340	3395	0.2171	89545708	34.30	58
	4	1	38	Matérn	1052	4206	-0.0618	192524506	25.89	50
		2	39	Circular	1908	7630	0.4761	65452585	34.23	65
		3	40	Circular	876	4381	0.5250	106857235	32.79	84
		4	41	Circular	629	5659	0.2603	103575352	34.67	64
	5	1	42	Spherical	342	3417	0.1736	199870272	43.01	78
		2	43	Circular	1090	7630	0.5219	73716461	35.25	44
		3	44	Circular	798	4788	0.5221	103316140	32.89	104
		4	45	Circular	885	5313	0.1645	250953969	50.10	37

Note: The clustering criteria include five different variables (area, age, bedrooms, bathrooms, and type of property). Clustering criterion 1 considers the first variable (area); clustering criterion 2, the first two variables (area and age); and so on. The best-modeled submarket is shown in bold, and its corresponding clustering criterion rows use grey background.

Table 4: Summarized Results of the Models Selected From Table 3.

	Number of Clusters	Lag width [m]	Cutoff distance [m]	Exponential [%]	Gaussian [%]	Spherical [%]	Circular [%]	Matérn [%]
All cases	45	1018	6418	6.7	15.6	11.1	**55.6**	11.1
Case A	10	867	6461	0	0	0	**80**	20
Case B	15	1219	7434	13.3	26.7	6.7	**40**	13.3
Case C	20	944	5634	5	15	20	**55**	5

In Case C (four submarkets), Table 4 shows that the circular model was the most common one (55%), followed by the spherical, Gaussian, exponential, and Matérn (20%, 15%, 5%, and

5%, respectively). Also, according to Table 3, the variogram that best described a submarket was obtained by applying clustering criterion 3. The best fit model for submarket 2 was the Gaussian one (with an R-Squared value of 0.6244, a cutoff distance of 4787 m, and a lag width of 532 m); for submarket 1, the Gaussian model (with an R-Squared value of -0.1942); for submarket 3, the circular model (with R-Squared values of 0.5565); and for submarket 4, the spherical model (with R-Squared values of 0.2171).

According to Table 3 for the grey highlighted clustering criteria, Case A showed only one of the two submarkets *properly* modeled (a relatively high R-Squared value). For Case B there was a relatively high R-Squared value, an intermediate value, and a low value for each of the three submarkets. For Case C, clustering criteria 3, yielded two submarkets *properly* modeled with the highest R-Squared values (0.62 and 0.56, respectively).

Submarkets and Clustering Criterion Effect

The effect of the number of submarkets is observed by considering the weighted average values for the R-squared and MAPE values for each of the three cases (Table 5). Since the weighted average includes poorly modeled submarkets, high MAPE and low R-squared values are obtained. The best behavior is found for Case B, with mean R-Squared and MAPE values of 0.3446 and 35.05, respectively.

The clustering criterion effect is observed by taking the weighted average of the five different clustering criteria (Table 6). When comparing clustering criterion 1 (area) and clustering criterion 2 (area and age), it can be noticed an improvement in the modeling as the R-Squared increases from 0.2720 to 0.3409. It means that the variable of age provides additional and valuable information for the model.

On the other hand, clustering criteria 3 and 4, present no substantial improvement compared to clustering criterion 2. A similar explanation could be that the number of bedrooms and bathrooms depends on the area variable. Finally, clustering criterion 5 presents the highest R-Squared value, which means that prediction results are improved if the model is provided with information about the type of housing unit (i.e., apartment or house).

Table 5: Weighted Average Values for Cases Obtained from Table 3

Case	**Mean R-Squared**	**Mean MAPE**[%]
Case A	0.3370	41.51
Case B	0.3446	35.05
Case C	0.3164	33.54

From this point on, we will present a complete analysis (as shown in Figure 3) of Case B (three submarkets) using clustering criterion 2 (area and age), as these two factors produce the best models according to the criteria discussed previously. Figure 4 shows the spatial distribution of the three different submarkets in Medellín, Colombia, in Case B.

Moreover, Table 7 provides a spatial description of the three submarkets defined applying clustering criterion 2 and includes the following information: submarket, number of data (N_D), data density, number of paired comparisons among data, percentage of paired

Table 6: Weighted Average Values for Cases Obtained from Table 3

Submarket	**Mean R-Squared**	**Mean MAPE**[%]
Submarket 1	0.2720	32.59
Submarket 2	0.3409	36.94
Submarket 3	0.3486	37.20
Submarket 4	0.3385	37.08
Submarket 5	0.3633	39.70

comparisons within a range of 1 km, distance within which 50% of the paired comparisons can be found, the mean and standard deviation of the distance among data, and mean and standard deviation of log price.

Submarkets 1, 2, and 3 have, respectively, 54, 65, and 144 observation points (N_P), with corresponding densities of 0.5376, 0.6471, and 1.4335 N_P/km^2. Submarket 3 (which is actually the best modeled) exhibits the highest mean distance (4979 m) because it includes most points, which are scattered all over the city. In turn, most observation points of submarkets 1 and 2 are found in Southeastern Medellín. Also, 3.65% of the paired comparisons of submarket 3 are located at a maximum distance of 1 km; and 50% of the connections, at a maximum distance of 11.51 km.

Figure 4: Map of Medellín Showing the Three Submarkets Defined Applying Clustering Criterion 2 (Area and Age)

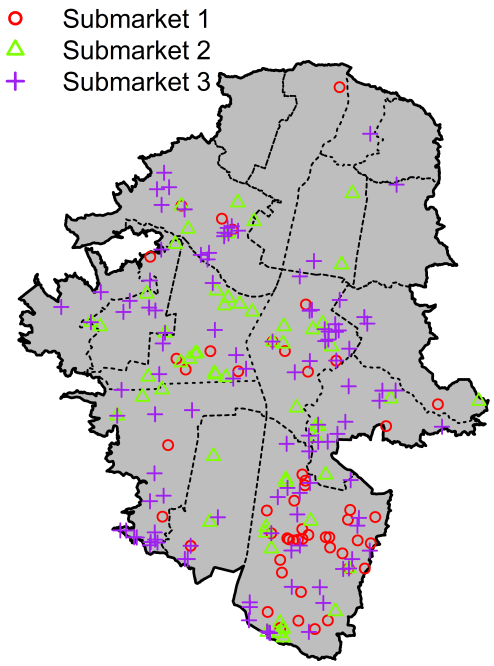


Table 7: Spatial Data Statistics of Case B (Three Submarkets) Using Clustering Criterion 2 (Area and Age)

Submarket	Number of Data (N_P)	Data Density (N_P/m^2)	Paired Comparisons	Percentage of Paired Comparisons Within 1 km	Distance Within Which 50 % of Paired Comparisons Can Be Found	Mean Distance (m)	SD of Distance (m)	Mean Log(Price) (COP)	SD of Log(Price) (COP)
1	54	0.5376	1431	4.40	8225	4041	2755	20.1481	0.36760714
2	65	0.6471	2080	3.65	8202	4515	2407	19.0574	0.54374478
3	144	1.4335	10296	3.65	11510	4979	2510	18.8644	0.70816505

Figure 5 shows selected semi-variograms of the three models analyzed in this study (i.e., Matérn, spherical, and circular) and the modeling parameters (nugget, range, and partial sill) for the three submarkets. Additionally, the validation indexes obtained via LOOCV are presented in Table 8, which is actually a subset of Table 3. Figure 5 indicates that the three submarkets are best modeled using large cutoff distance values and lag widths. Nevertheless, only submarket 3 yields proper results according to the R-Squared index.

Figure 5: Selected Semi-variogram Models for the Set with Three Submarkets Using Clustering Criterion 2 (Area and Age).

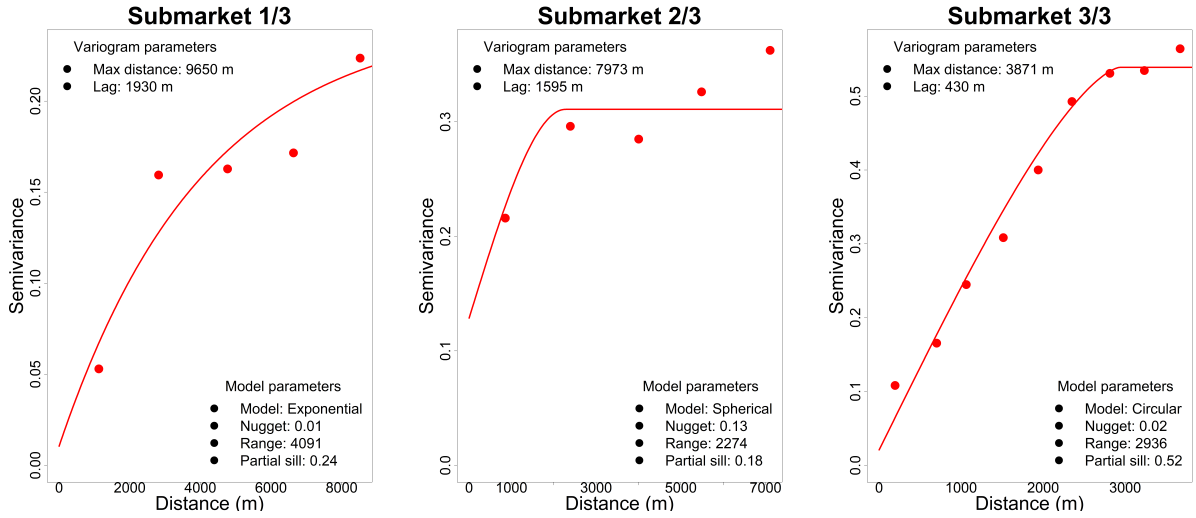


Table 8: Leave-One-Out Cross-Validation of the Model Selected for Case B (Set with Three Submarkets) Using Clustering Criterion 2 (Area and Age).

Submarket	Cluster ID	Model	Lag Width	Cutoff Distance	R-squared	RMSE	MAPE
1	14	Matérn	1930	9650	-0.1258	208438795	26.82
2	15	spherical	1594	7972	0.2201	86942896	42.34
3	**16**	**circular**	**1106**	**7743**	**0.5811**	**98650579**	**32.87**

The selected variograms were used to obtain the mean values on the residential property map, as shown in Figure 6 shows the interpolated kriging map (which is masked by the transport network) of each cluster in Medellín. This is an example of the isoline maps that can be obtained from kriging interpolation; in this case, it shows the results of 3 submarkets applying criterion 2 (area and age). In Table 8, the mean absolute percentage error (MAPE) of submarkets 1, 2, and 3 reached 26.82%, 42.34%, and 32.87%, respectively.

Figure 6: Map of Predicted Property Prices (Set With 3 Clusters) Obtained Applying Criterion 2 (Area and Age)

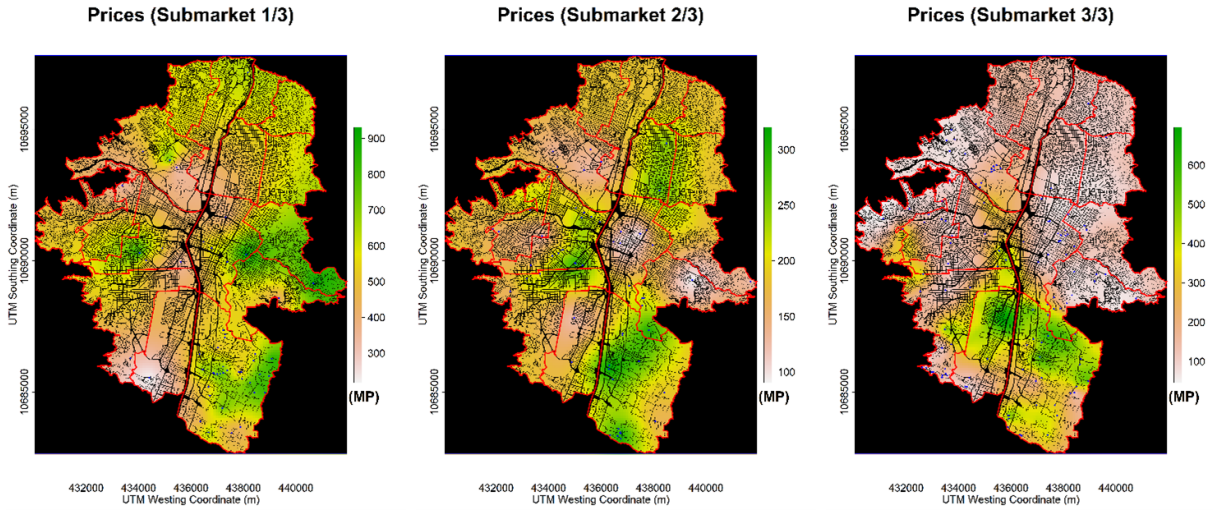


Figure 7 shows the Multidimensional Index of Living Conditions (MILC) according to the Encuesta de Calidad de Vida (Quality of Life Survey) [2] in the year 2018 of the 16 comunas (districts) of Medellín. This index serves as a global indicator of living conditions (on a scale from 1 to 100) for each district (Alcaldía de Medellín, 2019). According to the isoline maps of Figure 7, the southeast area of the city, which is district 14 (El Poblado), concentrates the most expensive housing. In fact, the MILC of district 14 is the highest.

The two-stage methodology approach offers interesting insights into the behavior of each submarket in relation to the living standards in the area of the dwellings. According to Figure 6, a dwelling of the submarkets 1 or 2 can cost up to three times more for being located in an area with a high MILC (such as in the southeast of the city), rather than being located in the northern side of the city (districts 1 to 7), where the MILC is under the average of the city (49.3). These differences in prices are much more noticeable for properties of submarket 3 (typically recently built and small apartments, according to Table 2). In this case, a dwelling can cost up to six times more for being located in the district with the highest MILC.

The predictions plotted in Figure 8 show the discrepancies in prices (log was applied for the sake of visualization) between Medellín Property Price Register and the values predicted for the third cluster, which is the best fit by a line with intercept 5.86 and slope 0.69. The fact that the dashed line (adjusted to the points) is above the solid line in the early phases of the chart indicates that, when property values are estimated, the kriging method overestimates low values and underestimates high ones.

[2]The Encuesta de Calidad de Vida (Quality of Life Survey) is an instrument whose purpose is to monitor and measure the socioeconomic conditions of the inhabitants of the 16 districts and 5 townships that make up the city of Medellín. It is a primary source of information that allows knowing indices on issues of marked importance such as population, housing, households, education, workforce, health, and social security.

Figure 7: Multidimensional Index of Living Conditions of the 16 Districts in Medellín (Alcaldía de Medellín, 2019)

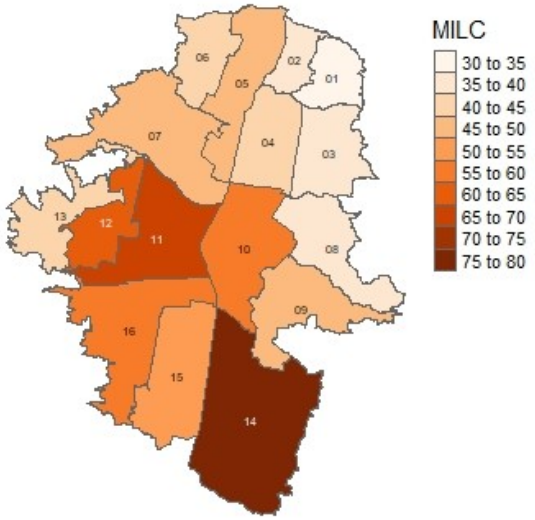
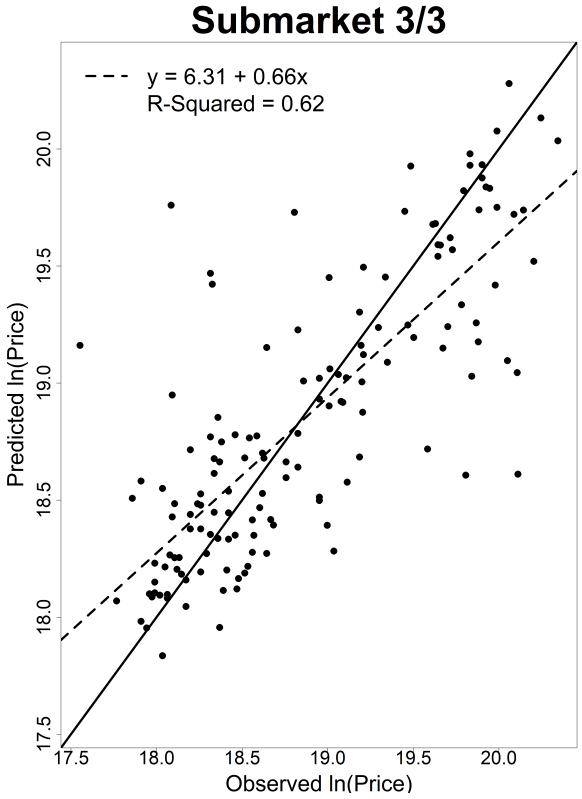


Figure 8: Observed vs. Predicted Property Prices (Set with 3 Clusters) Of Cluster 3 Applying Criterion 2 (Area and Age)



5. CONCLUSIONS AND RECOMMENDATIONS

In this study, we modeled house prices using a mixed two-stage model for mass appraisal and a sample of 293 second-hand house valuations conducted between 2014 and 2019 in Medellín,

Colombia. During the first stage, 2, 3, and 4 submarkets were generated by the k-means algorithm implementing 5 clustering criteria. The first criterion considered only the housing area; the second one the area and age of the property; and so on until criterion 5, which took into account the area, age, number of rooms, number of bathrooms, and type of property (house or apartment). During the second stage, by means of different kriging models, we incorporated spatial dependency into the estimation of house prices.

The results show that the number and geographic distribution of the housing units whose information is known are important to implement the model in practice. A higher number of submarkets implies lower number of housing units in each submarket. This implies larger areas for which there is no available information, increasing the uncertainty of the model Consequently, house price models that use clustering and kriging should strike a balance between the number of submarkets, the number of housing units per submarket, and their geographic distribution. Nevertheless, such balance depends directly on the size and geographic distribution of the housing units in the sample.

Property area and age turned out to be the best criterion to create submarkets, even better than a combination of the area and age with other characteristics (e.g., number of rooms, number of bathrooms, and type of property). This is due to the workings of the k-means clustering algorithm, which does not differentiate the importance of the variables or the ranges of the values they can take.

The best-fit kriging model was obtained using cross-validation, sweeping different parameters of the empirical variogram, which were adjusted by means of different theoretical models. In all the cases, it was shown that the most suitable model for the different clusters was circular, appearing 55.6% of the time, while its Gaussian, spherical and Matérn counterparts did so in proportions of 15.6%, 13.3%, and 11.1%, respectively. The least common model was exponential (4.4%). The mean Lag Width was 1064 m, and the cutoff distance, 6497 m.

The effect of the number of submarkets was studied in three cases with five clustering criteria by sweeping different parameters of the empirical variograms adjusted to specific theoretical models. In general, each clustering criterion (set of submarkets) produced one well-modeled submarket and other submarkets that achieved poor results. When considering the weighted average values for the R-squared and MAPE values for each of the three cases, the best behavior is found for Case B, with mean values of 0.3446 and 35.05% for R-Squared and MAPE, respectively.

The criterion effect was observed by taking the weighted average of the five different criteria. Compared to criterion 1 (area), criterion 2 (area and age) generates a noticeable improvement in the modeling as the R-Squared increases from 0.2720 to 0.3409. However, criteria 3 and 4 present no substantial improvement compared to criterion 2. Criterion 5 presents the highest R-Squared value.

An explanation of these results could be that variable age provides additional and valuable information for the model that only considers variable area. The number of bedrooms and bathrooms is dependent on the area, and that is why the model shows no improvement. Finally, providing the model the information about whether it is an apartment or a house can improve the prediction results.

The best results were obtained as follows. First, the sample was divided into three submarkets (Case B). Second, property area and age (criterion 2) were employed to classify the housing units. Then, a Matérn kriging model was applied to submarket 1; a spherical kriging model, to submarket 2; and a circular kriging model, to submarket 3. The MAPEs obtained for such submarkets were 26.82%, 42.24%, and 32.87%, respectively. These results are in line with Calka (2019), who indicated that "the minimum number of points in a cluster should not be below 30, but [...] to obtain an estimation error of less than 10% it should be around 200". In this case, the number of housing units in each cluster was 54, 65 and 144, respectively.

The two-stage methodology approach offers interesting insights into the behavior of each submarket in relation to the living standards in the area of the dwellings. A dwelling of the submarkets 1 or 2 can cost up to three times more for being located in an area with a high Multidimensional Index of Living Conditions (MILC) rather than being located in an area where the MILC is under the average of the city. These differences in prices are much more noticeable for properties of submarket 3 (typically recently built and small apartments) which can cost up to six times more for being located in an area with high MILC. This could mean that the prices of dwellings in submarket 3 benefit much more from public and private amenities than bigger and older constructions in submarkets 1 and 2.

The application of the methodology presented in this article could play an important role in the development of land use policies for the design of soil management policies and overall land planning. The difference in prices of a submarket made up mainly of recently built and small apartments calls the attention. This could indicate the existence of speculative practices that increase the price of this kind of properties. Therefore, new and small housing units should be under special control.

The relationship between submarkets, house prices and living standards can be an important tool in the identification of zones with a higher need for government interventions that promote more equitable land development. It also could help with more efficient funding via an improvement of the recouping of public investments by the state in the areas where the house prices benefit the most from public amenities.

Future studies of mixed house price models based on clustering and kriging should use classification algorithms other than the k-means to create the submarkets and include variables such as environmental pollution, health care, and education indexes.

REFERENCES

Alcaldía de Medellín. (2019) "Informe Calidad de Vida de Medellín 2018," Alcaldía de Medellín: Medellín, Colombia.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, 1st edition. http://doi.org/10.2307/143780.

Antipov, Evgeny A and Elena B Pokryshevskaya. (2012) "Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a Cart-Based Approach for Model Diagnostics," *Expert Systems with Applications*, 39(2), 1772–1778. http://doi.org/https://doi.org/10.1016/j.eswa.2011.08.077.

Appraisal Institute. (2013) *The Appraisal of Real Estate*. American Institute of Real Estate Appraisers, 14th edition.

Basu, Sabyasachi and Thomas G. Thibodeau. (1998) "Analysis of Spatial Autocorrelation in House Prices," *Journal of Real Estate Finance and Economics*, 17(1), 61–85. http://doi.org/10.1023/A:1007703229507.

Bishop, T F A and R M Lark. (2008) "A Comparison of Parametric and Non-parametric Methods for Modelling a Coregionalization," *Geoderma*, 148(1), 13–24. http://doi.org/https://doi.org/10.1016/j.geoderma.2008.08.010.

Bourassa, Steven C, Eva Cantoni, and Martin Hoesli. (2007) "Spatial Dependence, Housing Submarkets, and House Price Prediction," *The Journal of Real Estate Finance and Economics*, 35(2), 143–160. http://doi.org/10.1007/s11146-007-9036-8.

Calka, Beata. (2019) "Estimating Residential Property Values on the Basis of Clustering and Geostatistics," *Geosciences*, 9(3), 1–14. http://doi.org/10.3390/geosciences9030143.

Calka, Beata and Elzbieta Bielecka. (2016), "The Application of Geoinformation Theory in Housing Mass Appraisal," In *2016 Baltic Geodetic Congress (BGC Geomatics)*, pp. 239–243. https://www.sciencegate.app/document/10.1109/bgc.geomatics.2016.50.

Can, Ayse. (1992) "Specification and Estimation of Hedonic Housing Price Models," *Regional Science and Urban Economics*, 22(3), 453–474. http://doi.org/https://doi.org/10.1016/0166-0462(92)90039-4. Special Issue Space and Applied Econometrics.

Cebula, Richard J.. (2009) "The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District," *The Review of Regional Studies*, 39(1), 9–22.

Cellmer, Radoslaw, Mirosław Bełej, Sabia Zrobek, and Maruška Šubic Kovač. (2014) "Urban Land Value Maps - A Methodological Approach," *Geodetski Vestnik*, 58, 535–551.

Credit, Kevin. (2021) "Spatial Models or Random Forest? Evaluating the Use of Spatially Explicit Machine Learning Methods to Predict Employment Density Around New Transit Stations in Los Angeles," *Geographical Analysis*, 54, 58–83. http://doi.org/10.1111/gean.12273.

Córdoba, Mariano, Juan Pablo Carranza, Mario Piumetto, Federico Monzani, and Mónica Balzarini. (2021) "A Spatially Based Quantile Regression Forest Model for Mapping Rural Land Values," *Journal of Environmental Management*, 289, 112509. http://doi.org/https://doi.org/10.1016/j.jenvman.2021.112509.

DANE. (2019) "Censo Nacional. Información Capital," Departamento Administrativo Nacional de Estadísticas (DANE): Bogotá, Colombia.

de Marsily, G.. (1984) "Spatial Variability of Properties in Porous Media: A Stochastic

Approach," *Fundamentals of Transport Phenomena in Porous Media*, 82, 719–769. http://doi.org/10.1007/978-94-009-6175-3_15.

Gabrielli, Laura, Salvatore Giuffrida, and Maria Rosa Trovato. (2017), "Gaps and Overlaps of Urban Housing Sub-market: Hard Clustering and Fuzzy Clustering Approaches," In Stanghellini, Stefano, Pierluigi Morano, Marta Bottero, and Alessandra Oppio, eds., *Appraisal: From Theory to Practice: Results of SIEV 2015*, pp. 203–219, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-49676-4$_1$5.

García, Noelia, Matías Gámez, and Esteban Alfaro. (2008) "ANN+GIS: An Automated System for Property Valuation," *Neurocomput*, 71(4–6), 733–742. http://doi.org/10.1016/j.neucom.2007.07.031.

Hengl, Tomislav. (2009) "A Practical Guide to Geostatistical Mapping (Vol. 52)," *Amsterdam, The Netherlands: University of Amsterdam*.

Johnston, Kevin, Jay M Ver Hoef, Konstantin Krivoruchko, and Neil Lucas. (2001) *Using Arcgis Geostatistical Analyst*, volume 380. Esri Redlands.

Lai, Peddy Pi-Ying. (2011) "Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City," *Journal of Modern Accounting and Auditing*, 7, 1080–1089.

Lark, R M. (2000) "A Comparison of Some Robust Estimators of the Variogram for Use in Soil Survey," *European Journal of Soil Science*, 51(1), 137–157. http://doi.org/https://doi.org/10.1046/j.1365-2389.2000.00280.x.

Ling, David C. and Wayne R. Archer. (2018) *Real Estate Principles*. McGraw-Hill Higuer Education, 5th edition. http://doi.org/10.2307/1233122.

Lozano-Gracia, Nancy and Luc Anselin. (2012) "Is the Price Right?: Assessing Estimates of Cadastral Values for Bogotá, Colombia," *Regional Science Policy & Practice*, 4(4), 495–508. http://doi.org/https://doi.org/10.1111/j.1757-7802.2012.01062.x.

Marko, Kuswantoro, Nassir S Al-Amri, and Amro M M Elfeki. (2014) "Geostatistical Analysis Using GIS for Mapping Groundwater Quality: Case Study in the Recharge Area of Wadi Usfan, Western Saudi Arabia," *Arabian Journal of Geosciences*, 7(12), 5239–5252. http://doi.org/10.1007/s12517-013-1156-2.

McCluskey, William, Dzurllkanian Zulkarnain Daud, and Norhaya Kamarudin. (2014) "Boosted Regression Trees: An Application for the Mass Appraisal of Residential Property in Malaysia," *Journal of Financial Management of Property and Construction*, 19(2), 152–167. http://doi.org/10.1108/JFMPC-06-2013-0022.

McCluskey, William, Peadar Davis, Martin Haran, Michael McCord, and David McIlhatton. (2012) "The Potential of Artificial Neural Networks in Mass Appraisal: The Case Revisited," *Journal of Financial Management of Property and Construction*, 17(3), 274–292. http://doi.org/10.1108/13664381211274371.

Mehrjardi, R. Taghizadeh, M Zareian Jahromi, Sh Mahmodi, and A Heidari. (2008) "Spatial Distribution of Groundwater Quality with Geostatistics (Case Study: Yazd-Ardakan Plain)," *World Applied Sciences Journal*, 4(1), 9–17.

Mimis, Angelos, Antonis Rovolis, and Marianthi Stamou. (2013) "Property Valuation With Artificial Neural Network: The Case of Athens," *Journal of Property Research*, 30(2), 128–143. http://doi.org/10.1080/09599916.2012.755558.

Minasny, Budiman and Alex B McBratney. (2005) "The Matérn Function as a General Model for Soil Variograms," *Geoderma*, 128(3-4), 192–207.

Monson, Matt. (2009) "Valuation Using Hedonic Pricing Models," *Cornell Real Estate Review*,

7, 10.

Montero, José María, Román Mínguez, and Gema Fernández-Avilés. (2018) "Housing Price Prediction: Parametric Versus Semi-parametric Spatial Hedonic Models," *Journal of Geographical Systems*, 20(1), 27–55. http://doi.org/10.1007/s10109-017-0257-y.

Morano, Pierluigi, Francesco Tajani, and Marco Locurcio. (2018) "Multicriteria Analysis and Genetic Algorithms for Mass Appraisals in the Italian Property Market," *International Journal of Housing Markets and Analysis*, 11(2), 229–262. http://doi.org/10.1108/IJHMA-04-2017-0034.

Napoli, Grazia, Salvatore Giuffrida, and Alberto Valenti. (2017), "Forms and Functions of the Real Estate Market of Palermo (Italy). Science and Knowledge in the Cluster Analysis Approach," In Stanghellini, Stefano, Pierluigi Morano, Marta Bottero, and Alessandra Oppio, eds., *Appraisal: From Theory to Practice: Results of SIEV 2015*, pp. 191–202, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-49676-$4_1$4.

Nayanaka, V, W Vitharana, and R Mapa. (2011) "Geostatistical Analysis of Soil Properties to Support Spatial Sampling in a Paddy Growing Alfisol," *Tropical Agricultural Research*, 22(1), 34–44. http://doi.org/http://doi.org/10.4038/tar.v22i1.2668.

Reyes-Bueno, Fabián, Juan Manuel García-Samaniego, and Aminael Sánchez-Rodríguez. (2018) "Large-Scale Simultaneous Market Segment Definition and Mass Appraisal Using Decision Tree Learning for Fiscal Purposes," *Land Use Policy*, 79, 116–122. http://doi.org/https://doi.org/10.1016/j.landusepol.2018.08.012.

Selim, Hasan. (2009) "Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network," *Expert Systems with Applications*, 36(2, Part 2), 2843–2852. http://doi.org/https://doi.org/10.1016/j.eswa.2008.01.044.

Stevens, Benjamin H. (1968) "Location Theory and Programming Models: The von Thünen Case," *Papers of the Regional Science Association*, 21(1), 19–34. http://doi.org/10.1007/BF01952719.

Teixeira, M.C.C., J.M. Caridad, and N. Ceular. (2010), "Hedonic Methodologies in the Real Estate Valuation," In *Mathematical Methods in Engineering International Symposium, Coimbra, Portugal, October 2010*. Instituto Politécnico de Coimbra. http://hdl.handle.net/10400.11/412.

Vattani, Andrea. (2011) "K-Means Requires Exponentially Many Iterations Even in the Plane," *Discrete & Computational Geometry*, 45(4), 596–616. http://doi.org/10.1007/s00454-011-9340-1.

Čeh, Marjan, Milan Kilibarda, Anka Lisec, and Branislav Bajat. (2018) "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," *ISPRS International Journal of Geo-Information*, 7(5), 1–16. http://doi.org/10.3390/ijgi7050168.

Vo, Nguyen, Hao Shi, and Jukab Szajman. (2015) "Sensitivity Analysis and Optimisation to Input Variables Using Wingamma and Ann: A Case Study in Automated Residential Property Valuation," *International Journal of Advanced and Applied Sciences*, 2(12), 19–24.

Wang, Daikun and Victor Jing Li. (2019) "Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review," *Sustainability*, 11(24). http://doi.org/10.3390/su11247006.

Ward, Joe H. (1963) "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58(301), 236–244. http://doi.org/10.1080/01621459.

1963.10500845.