VALORACIÓN MASIVA DE INMUEBLES RESIDENCIALES MEDIANTE MODELOS MULTINIVEL

MASS APPRAISAL OF RESIDENTIAL REAL ESTATE USING MULTILEVEL MODELLING

Iván ARRIBAS
Universidad de Valencia, Facultad de Ciencias Económicas; ERI-CES; Ivie, Spain
Email: ivan.arribas@uv.es

Rima TAMOŠIŪNIENĖ
Vilnius Gediminas Technical University ; Mykolas Romeris University, Lithuania
Email: rima.tamosiuniene@vgtu.lt

Resumen:
La correcta valoración de los inmuebles inmobiliarios, y en especial de las viviendas, es un tópico de gran importancia económica. Dentro de los métodos de valoración masiva de viviendas, los modelos econométricos son los que se emplean con mayor asiduidad. No obstante, estos modelos presentan algunas deficiencias dado que asumen la hipótesis poco realista de que los individuos de la muestra son independientes entre sí y no consideran la estructura jerárquica de los datos. En este trabajo se utiliza el modelo lineal jerárquico (HLM por sus siglas en inglés) para subsanar esta deficiencia y se aplica a una base de datos estructurada jerárquicamente en tres niveles: vivienda; barrio y ciudad. Se comprueba cómo el modelo obtenido mejora al modelo econométrico tradicional tanto en la bondad del ajuste como en términos de varianzas estimadas.

Palabras clave: Mercado inmobiliario residencial, modeles jerárquicos, modelos multinivel, valoración masiva

Abstract

Accurate real estate valuation, especially residential real estate valuation, is a topic with a high social and economic impact. Among the methods employed for mass valuation, econometric models are the more widely used. Nevertheless, these models have some drawbacks as they assume some hypothesis which are unrealistic, considering the individuals in the sample as independent and they do not consider the hierarchical data structure. This paper applies a Hierarchical Lineal Model (HLM) to overcome this problem. The model is employed on a big database with three hierarchical levels: apartment, neighborhood and city. It is concluded that HLM improves the traditional econometric model OLS both in terms of goodness of fit and residual variance.

Keywords: Residential real estate market; hierarchical models; multilevel models, mass appraisal

JEL: C5,D4,R3

1. INTRODUCTION

Real estate market has an undoubtable importance on the economy, having a great impact on the employment, financial system, public policies and family expenditure, among many other economic variables. This fact has become even more obvious during the recent economic and financial crisis. As a result, it is not surprising that the real estate market has been analyzed from many different perspectives such as price formation (Aznar *et al.*, 2010), market liquidity (Zeng *et al.,* 2013; Pestana Barros *et al.*, 2014), price evolution (I-Chun, 2014; Feng and Wu, 215; Plakandaras *et al.*, 2015), or decision making in real estate investment (Cervelló *et al.*, 2011), just to mention a few research topics.

Another topic that has attracted the attention of academics and practitioners is real estate mass valuation. Following Gloudemans (1999), mass appraisal is the systematic appraisal of groups of properties as of a given date using standardized procedures and statistical testing. Mass appraisal is characterized by the intensive use of standardized procedures on a large database, including high number of observations and explanatory variables of the price. As for residential real estate valuation, there are many organizations, both public and private, which take advantage of mass appraisal methodologies. For example, valuation companies used them to control the quality of valuations made by professional appraisers; banks employ mass appraisals to value real estate employed to issue mortgage backed securities, real estate investment funds calculate the value of real estate portfolios, local governments use it to calculate taxes etc.

The need for mass appraises has triggered the development of different methodologies based, on the one hand, on the econometrics and, on the other hand, on artificial intelligence. The econometric approach includes all kind of regression models. Artificial intelligence has introduced new solutions like decision trees (Fanet *et al.*, 2006), rough set theory (d'Amato, 2007), artificial neural networks (Tay and Ho, 1991; García *et al.*, 2008; Selim, 2009), support vector machines (Kontrimas and Verikas, 2011) and random forest (Antipov and Pokryshevskaya, 2012).

This paper applies an econometric model, the hierarchical linear model (hereafter HLM), also known as multilevel model, mixed model, random effects model and variance components model on real estate mass valuation. This methodology is already popular in other fields, but its use in residential real estate mass valuation is rather scarce.

In fact, hierarchical linear models have been successfully applied since the 80's in various fields such as education (Aitkin *et al.*, 1981; Raudenbush and Bryck, 1986; Singh, 2014), public policy (Duncan *et al.*, 1993; Tso and Guan, 2014), criminology (Gelman, 2007; Fagan *et al.*, 2015), and politics (Wang *et al.*, 2014). These models overcome some limitations of the traditional regression models, which rely on the hypothesis that the individuals in the sample are independent, although this is not necessarily the case. Furthermore, HLM presents valuable additional information regarding the percentage of the variance error generated by each of the level in the hierarchical structure. In this paper, variables are grouped into three hierarchy levels, so the role of variables in the apartment, neighborhood and city level can be determined.

Together with this three-level structure of the data, another important contribution of this research is the use of a vast database containing many observations and many explanatory variables. Therefore, the results obtained are more robust than other similar studies, which employed a more limited set of data in terms of number of observations and variables.

The remainder of the paper is structured as follows. Next section presents previous studies that applied econometric models to residential real estate mass valuation. Section 3 introduces the Hierarchical Linear Model (HLM), section 4 is devoted to the description of the database use to estimate the econometric models. Section 5 shows the results obtained when applying HLM and OLS on residential real estate mass valuation in the three most populated cities in Spain: Madrid, Barcelona and Valencia. Finally, section 6 concludes.

2. MASS APPRAISAL OF RESIDENTIAL REAL ESTATE USING ECONOMETRICS MODELS

Econometric models, more specifically hedonic regression models, have been applied in many academic research studies and are widely used among practitioners. In fact, many papers have been published since the 80’s which employ econometric models of different complexity such as the traditional hedonic regression models (Palmquist, 1984; Isakson, 2001; Downes and Zabel, 2002), ridge regression (Ferreira and Sirmans, 1988) or quantile regression (Farmer and Lipscomb, 2010; Narula *et al.*, 2012) etc.

Hedonic models are used to determine which factors influence the price of the appartments, including both characteristics of the flats and of the environment, as well. Nevertheless, traditional hedonic models assume that these characteristics are independent, implicitly considering that the characteristics of the flat and their impact on the price are constant, regardless the location of the dwelling. This premise does not reflect the reality, as neighborhood characteristics are not independent of the dwelling and both neighborhood and dwelling influence each other.

The problem is that when spatial autocorrelation exists in the error term in a hedonic price equation, the assessment results of the parameters may be subject to error (Basu and Thibodeau, 1998). Incorrect coefficients may also be caused by the explanatory variables in the model, leading to wrong conclusions.

Considering that the impact on the price of the characteristics of the flat is independent of the location of the flat, means that the multilevel or hierarchical structure that generates the prices of the dwellings is not considered. But this hierarchical structure does really exist, although it is not captured by the hedonic models, as the flats are located in neighborhoods and neighborhoods are located in cities.

Brown and Uyar (2004) consider that HLM can be applied to overcome these problems and correctly assess the implicit price of a house with non-constant variance and spatial heterogeneity. In other words, HLM can be used to separate the variation in housing prices into a portion that depends on house-specific characteristics and another portion that depends on neighborhood-specific characteristics. These authors also note that although GIS data can be used for neighborhood effects and spatial correlation, it will not identify the impact of individual neighborhood characteristics on the price of a house.

Even though the use of HLM on massive real estate valuation is very promising, only few studies can be found which apply this methodology. Among those pioneer papers we can mention Jones and Bullen (1993), who introduced HLM with two levels, but using a very reduced simple; Goodman and Thibodeau (1998), which includes information about the quality of the schools within the town; and Orford (2000), which applies a multilevel model to analyse the spatial structures in the residential real estate market and their impact on valuation. In this century, Brown and Ullar (2004) applied HLM to examine the influence of the characteristics of the flat and the neighborhood on the price, but they only use the area as the only variable to describe the dwelling. Another interesting paper is Lee (2009), which analyses the influence of satisfaction with public facilities on housing prices. Finally other important research papers are those by Giuliano et al. (2010), which studies the relationship between accessibility and residential land value; Chasco and Le Gallo (2013) which relate the price of the flat with air quality and noise, and Glaeser and Caruso (2015) about the role of space diversity and the mix of neighborhood services on the price of residential land in Luxembourg.

Those studies applying HLM on residential real estate mass valuation are even scarcer. Only in recent times some papers have been published about this topic, which analyse the housing submarkets within a city (Leishman *et al.*, 2013; Arribas *et al.*, 2016) or at an administrative- district level (Lee and Lin, 2014).

To our knowledge, this is the first study that applies a three-level hierarchical linear model to perform mass appraisal of residential real estate in Spain using a vast database and variables to describe apartment, block and neighborhood characteristics, that is, the information employed by practitioners in Spain to make valuations following the instructions and regulations issued by the Banco de España.

3. METHODOLOGY

Our analysis applies HLM, as the database is hierarchically structured with apartments (Level 1 units) geographically allocated in postal codes (Level 2 units) taken from three Spanish cities (Level 3 unit), those with the highest population. Postal codes are used as a proxy for the different neighborhoods in a city.

HLM models are statistically more efficient than analyses that only consider the apartment level, those that only consider the postal code level, or those that use both with data panel techniques. As mentioned above, considering the clustered nature of the database allows unbiased effects and robust standard errors to be estimated and correct significance tests to be produced. HLM also allows for the inclusion of variables at apartment level (e.g. number of rooms), postal code level (e.g. commercial characteristics of the neighborhood), or city level (e.g. population).

HLM was fitted to the data using the *xtmixed* command in Stata statistical software version 12.0 (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP). Different models were fitted for different levels for explanatory variables. The output of these analyses has two parts: *fixed effects* for each explanatory variable that are interpreted in the ordinary multiple regression sense (the average effect of the explanatory variable on the response variable); and *random effects* that describe the unexplained variability in the response variable. There are three random parameters, one for the Level 1 (apartment) variation, one for Level 2 (postal code) and one for the Level 3 (city) variation. This comparison makes it possible to estimate the percentage of variation attributable to Levels 2 and 3 (variance partition coefficients).

The statistical significance of any of the estimated parameters is tested by comparing the goodness of fit of two alternative models and testing whether the improvement in fit is statistically significant.

3.1. Multilevel model formulation

The literature on residential real estate valuations has considered a variety of alternative models in which the set of explanatory variables depends partially on the information available: size of the apartment, number and type of rooms, year of construction, characteristics of the block and the floor on which the apartment is located, among others. The most frequently used dependent variables are *apartment price* and *price per square meter*. Given the hierarchical nature of the data in which apartments are nested within geographical areas (postal codes), which are also nested within cities, the three-level HLM is written as

$$Y_{ijk} = \beta_{0jk} + \boldsymbol{\beta}'\boldsymbol{x} + e_{ijk}, \qquad i = 1, \dots, N; j = 1, \dots, J; k = 1, \dots, K \qquad [1]$$

$$\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk}, \qquad j = 1, \dots, J; k = 1, \dots, K \qquad [2]$$

$$v_{0k} \sim N(0, \sigma_{v0}^2) \qquad [3]$$

$$u_{0jk} \sim N(0, \sigma_{u0}^2) \qquad [4]$$

$$e_{ijk} \sim N(0, \sigma_e^2) \qquad [5]$$

where Y_{ijk}represents the price for apartment *i* in postal code *j* in city *k*; v_{0k} is the random effect of city *k*, and u_{0jk} is the random effect of postal code *j*, within city *k*. In a fixed effects model both effects the postal code and the city, represented by $\beta_0 +$

$\beta_k + \beta_{jk}$, are treated as fixed parameters for which direct estimates are obtained. In a HLM (random effect model), the apartment effects are independent random variables, whose distributions are summarized by two parameters, the mean (zero) and variances σ_{v0}^2 and σ_{u0}^2. Thus, the combined random effect is represented as $\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk}$, where β_0 is the constant mean price across apartments. Moreover, e_{ijk} is the residual error term, which is independent of the random effects, and has zero mean and variance σ_e^2. Vector $\boldsymbol{x}$ represents the explanatory variables at any level and $\boldsymbol{\beta}$ their estimated effect.

Notice that price variance for apartment *ijk* (total variance) is defined as

$$var(Y_{ijk}) = var(v_{0k} + u_{0jk} + e_{ijk}) = var(v_{0k}) + var(u_{0jk}) + var(e_{ijk})$$
$$= \sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2.$$

However, the covariance of two apartments depends on their allocation. The three following covariances correspond to two apartments in i) different cities (and also different postal codes), ii) the same city but different postal codes, and iii) same city and postal code, respectively:

i) $covar(Y_{ijk}, Y_{i'j'k'}) = covar(v_{0k} + u_{0jk} + e_{ijk}, v_{0k'} + u_{0j'k'} + e_{i'j'k'}) = covar(v_{0k}, v_{0k'}) + covar(u_{0jk}, u_{0j'k'}) + covar(+e_{ijk}, e_{i'j'k'}) = 0$

$i \neq i', j \neq j', k \neq k'.$

ii) $covar(Y_{ijk}, Y_{i'j'k}) = covar(v_{0k} + u_{0jk} + e_{ijk}, v_{0k} + u_{0j'k} + e_{i'j'k}) = covar(v_{0k}, v_{0k}) + covar(u_{0jk}, u_{0j'k}) + covar(+e_{ijk}, e_{i'j'k}) = \sigma_{v0}^2$

$i \neq i', j \neq j'.$

iii) $covar(Y_{ijk}, Y_{i'jk}) = covar(v_{0k} + u_{0jk} + e_{ijk}, v_{0k} + u_{0jk} + e_{i'jk}) = covar(v_{0k}, v_{0k}) + covar(u_{0jk}, u_{0jk}) + covar(+e_{ijk}, e_{i'jk}) = \sigma_{v0}^2 + \sigma_{u0}^2$

$i \neq i'.$

For example, the following 6×6 covariance matrix is derived from a sample of six apartments, four of them (apartments 1 to 4) allocated in the same city and the other two (apartments 5 to 6) allocated in a second city. Moreover, apartments 1 and 2 are grouped in the same postal code, apartments 3 and 4 are grouped in the same postal code but different of the previous postal code, and apartments 5 and 6 are in different postal codes.

$$\begin{pmatrix} \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 & \sigma_{v0}^2+\sigma_{u0}^2 & \sigma_{v0}^2 & \sigma_{v0}^2 & 0 & 0 \\ \sigma_{v0}^2+\sigma_{u0}^2 & \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 & \sigma_{v0}^2 & \sigma_{v0}^2 & 0 & 0 \\ \sigma_{v0}^2 & \sigma_{v0}^2 & \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 & \sigma_{v0}^2+\sigma_{u0}^2 & 0 & 0 \\ \sigma_{v0}^2 & \sigma_{v0}^2 & \sigma_{v0}^2+\sigma_{u0}^2 & \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 & \sigma_{v0}^2 \\ 0 & 0 & 0 & 0 & \sigma_{v0}^2 & \sigma_{v0}^2+\sigma_{u0}^2+\sigma_e^2 \end{pmatrix}$$

Model [1]-[5] can be estimated by ordinary least squares (OLS), assuming that the apartments under study are independent. More specifically, OLS assumes that the unique random effect is the residual term e_{ijk}, which is uncorrelated across postal codes and cities, $Var(\boldsymbol{e}) = \sigma^2 I$. However, in grouped data the group effect must be considered, which means that the independence assumption will not hold, (see the previous covariance matrix). One effect of ignoring clustering is that the standard errors of OLS parameters will be incorrectly estimated and no inference can be properly applied. HLM can estimate the correct standard errors and 99nalyse the nature of between-group variability and the effect of a grouping-level characteristic on an individual outcome, identify outlying groups and estimate group effects simultaneously with the effects of group-level explanatory variables.

Under HLM, total variance for individual *ijk* is $\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2$. The proportion of total variance explained by each level of the model is called the *variance partition coefficient* (VPC). Thus, the city level VPC is calculated as the ratio between the city variance to the total variance$VPC_v = \sigma_{v0}^2/(\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2)$; while the postal code level VPC is calculated as the ratio between the postal code variance to the total variance $VPC_u = \sigma_{u0}^2/(\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2)$; finally, the apartment VPC is $VPC_e = \sigma_e^2/(\sigma_{v0}^2 + \sigma_{u0}^2 + \sigma_e^2)$

Model [1]-[5] assume fixed coefficients for the explanatory variables, i.e., the slope coefficients of those explanatory variables are assumed fixed across higher-level units.

However, it could be the case that the relationship between the price and an explanatory variable varies across level 3 units (city) or level 2 units (postal code), or both. The HLM that allow the effects of explanatory variables to vary randomly across higher-level units are called random coefficients or random slope models. Let us see an example of a three-level random coefficient model where the coefficient of explanatory variable x_{1ijk} is random at both level two and at level three, and variable x_{2jk} is random only at level three. The model is written as:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}\, x_{1ijk} + \beta_{2k}\, x_{2jk} + e_{ijk}, \quad i = 1, \dots, N; j = 1, \dots, J; k = 1, \dots, K \qquad [6]$$

$$\beta_{0jk} = \beta_0 + v_{0k} + u_{0jk}, \quad j = 1, \dots, J; k = 1, \dots, K \qquad [7]$$

$$\beta_{1jk} = \beta_1 + v_{1k} + u_{1jk}, \quad j = 1, \dots, J; k = 1, \dots, K \qquad [8]$$

$$\beta_{2k} = \beta_2 + v_{2k}, \quad k = 1, \dots, K \qquad [9]$$

$$\begin{pmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{v0}^2 & & \\ \sigma_{v01} & \sigma_{v1}^2 & \\ \sigma_{v02} & \sigma_{v12} & \sigma_{v2}^2 \end{pmatrix} \right\} \qquad [10]$$

$$\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix} \right\} \qquad [11]$$

$$e_{ijk} \sim N(0, \sigma_e^2) \qquad [12]$$

The new random coefficients in the model are v_{1k}, which represents the random slope of variable x_{1ijk} at level three; u_{1jk}, which represents the random slope of variable x_{1ijk} at level two; and v_{2k}, which represents the random slope of variable x_{2jk} at level three. Now, the three level 3 random effects (one for the constant of the model, and two for the slopes of x_{1ijk} and x_{2jk}) are assumed to distribute as trivariate normal with zero mean and a 3×3 covariance matrix. The two level 2 random effects (one for the constant and one for the slope of variable x_{1ijk}) are assumed bivariate normal with zero mean and a2×2 covariance matrix. Random effects from different levels are independent, but the random effects from the same level could be correlated.

In random coefficients models, the between-group variance at each level is a function of the variables made random at that level. Thus, for model [6]-[12] the level 3 variance is a function of x_{1ijk} and x_{2jk}; the level 2 variance is a function of x_{1ijk}; and the level 1 variance is constant:

- Level 3 variance: $var(v_{0k} + v_{1k}x_{1ijk} + v_{2k}x_{2jk}) =$

$$\sigma_{v0}^2 + 2\sigma_{v01}x_{1ijk} + 2\sigma_{v02}x_{2jk} + \sigma_{v1}^2 x_{1ijk}^2 + 2\sigma_{v12}x_{1ijk}x_{2jk} + \sigma_{v2}^2 x_{2jk}^2.$$

- Level 2 variance: $var(u_{0jk} + u_{1jk}x_{1ijk}) = \sigma_{u0}^2 + 2\sigma_{u01}x_{1ijk} + \sigma_{u1}^2 x_{1ijk}^2.$
- Level 1 variance: $var(e_{ijk}) = \sigma_e^2.$

Thus, VPC is also a function of x_{1ijk} and x_{2jk}, and the way to interpret those variances is by plotting them against the explanatory variables.

4. DATABASE

Together with the application of HLM to mass valuation, the other novel feature of the present study is the large database employed, which consists of information on 8,685 apartments in the cities of Barcelona (1,235 apartments), Madrid (2,861 apartments) and Valencia (4,589 apartments), Spain. 16 variables were collected for each apartment: 6 describe its characteristics (Level 1 effect), 6 describe the building in which it is sited (Level 1 effect), and 3 describe the neighborhood (Level 2 effect). We have also considered the number of mortgages in each city, a Level 3 variable.

This large database contrasts with those used in most of the studies in this field, regardless of the valuation method applied, e.g. Brown et al., (2004) considered data for 725 dwellings and only one explanatory variable; d'Amato (2007) worked with 390 observations; García et al. (2008) used 591 sample cases; in Kontrimas et al., (2011) the sample size was 100; and Narula *et al.*, (2012) considered 54 observations.

No variable was recorded that had either almost empty levels or too many levels. The aim is to reduce the number of levels and thus significantly reduce the number of parameters to be estimated in OLS and HLM. There are no missing values in the database because real estate appraisers need all this information to assess the property. Each apartment is described by the following variables:

A. Apartment characteristics

a.1. Price: apartment price in euros

a.2. Area: total area of the apartment in square meters

a.3. Terrace: binary variable indicating whether or not the apartment has a terrace

a.4. Floor: floor on which the apartment is located

a.5. Bedrooms: number of bedrooms

a.6. Bathrooms: number of bathrooms

B. Block characteristics, which include both quantitative and qualitative variables
b.1. Number of apart.: number of apartments in the block
b.2. Lifts: binary variable indicating whether or not the block has a lift
b.3. Age of block: age of the block in years
b.4. Location: indicates the position of the block as a qualitative variable in four levels. "Very good" means that the building is near the sea-front or an important facility. "Good" is assigned when it is in a boulevard or large square. "Fair" is for an average street or thorough fare and "Bad" means it is in a narrow street or poor neighborhood.
b.5.Quality: describes the quality of the block construction as a qualitative variable. Three levels are considered: "High", "Medium" and "Low".
b.6. Community spaces: this qualitative variable indicates the existence of community spaces. Buildings are clustered into three groups: "None"(no community spaces),"Green areas" and "Green areas and more community spaces" (such as swimming pool, tennis court, etc.).

C. Neighborhood characteristics, which only include qualitative variables:
c.1. Commerce: commercial activity. Can be described as "Bad", "Regular". "Good" or "Very good".
c.2. Neighborhood: general perception of the neighborhood. Can be "Bad/Fair", "Good" or "Very good".
c.3. Income: perception of the neighborhood residents' income group classified into "High", "Medium-High", and "Medium-Low".
We have also considered the number of mortgages in each city, a level 3 variable. We could expect that prices and mortgages caused to each other. Thus, to avoid this endogeneity issue we have use the number of mortgages of the previous year.
Descriptive statistics for both quantitative and qualitative variables are given in Tables 1-3. The representative apartment in the whole sample (single-family residential property) has 93.6 square meters, no terrace, 2.8 bedrooms,1.4 bathrooms and is worth €275,500 (mean values). There are differences among cities. Barcelona is more expensive than the others, but Valencia has the biggest apartments, with also more bedrooms and bathrooms. In Madrid the apartments are, in average, the smallest but with more rooms than those in Barcelona.

Price is extremely skewed to the right, as shown by standard deviation. We partially solve this problem considering the log of price. Following most previous studies, we also considered the log transformation for several variables: Area, Floor, Bedrooms, Bathrooms, and Number of apartments.

Table 1: Descriptive statistics for quantitative variables

	Total		Barcelona		Madrid		Valencia	
	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation	Mean	Std. deviation
Apartment								
Price	275,485.1	181,297.6	377,030.5	212,890.5	310,339.2	197,754.6	226,427.2	140,373.3
Area	93.6	32.0	92.6	33.3	81.8	33.3	101.3	28.4
Terrace (*)	0.1	0.3	0.2	0.4	0.1	0.3	0.1	0.3
Floor	4.1	2.4	3.2	2.3	3.9	2.3	4.5	2.3
Bedrooms	2.8	0.9	2.7	0.9	2.4	0.9	3.0	0.8
Bathrooms	1.4	0.6	1.4	0.6	1.3	0.6	1.5	0.5
Block								
Lifts (*)	0.7	0.5	0.8	0.4	0.6	0.5	0.7	0.4
Number of apart.	21.6	16.0	21.9	15.7	21.5	16.2	21.7	15.9
Number of floors	6.8	2.6	6.3	2.6	6.2	2.5	7.3	2.6
Age of block	15.2	10.6	11.5	6.8	13.7	8.8	17.1	12.1

(*) 1 = Yes; 0 = No.

The typical block has 21.6 apartments, a lift, 6.8 floors, is 15 years old, has no community spaces, construction quality is “Medium” and its location is defined as “Good” (Tables 1 and 2). Most apartments are in a “Good” neighborhood, with “Medium-Low” income and “Good” commercial services (Table 3). Apartments in Barcelona are in average in worst blocks and worst neighborhoods than the ones in other cities, being the best ones in Valencia.

The lowest level of the qualitative variables that describe the block/neighborhood was chosen as the reference level in the analysis. For example, the reference level for the commercial services in the regressions is Bad to simplify the interpretation of the coefficient signs in the regression models.

The number of mortgages was 204,241 for Barcelona; 203,726 for Madrid; and 120,308 for Valencia.

Table 2: Description of the real estate qualitative variables

	Total		Barcelona		Madrid		Valencia	
	N	%	N	%	N	%	N	%
Location								
Very good	213	2.4	19	1.5	138	4.8	56	1.2
Good	7,555	87.0	992	80.3	2,435	85.1	4,128	90.0
Fair	787	9.1	196	15.9	264	9.2	327	7.1
Bad	130	1.5	28	2.3	24	0.8	78	1.7
Quality								
High	500	5.8	21	1,7	159	5.6	320	7.0
Medium	6,339	73.0	1,166	94.4	2,115	73.9	3,058	66.6
Low	1,846	21.3	48	3.9	587	20.5	1,211	26.4
Community spaces								
Green areas and more	713	8.2	46	3.7	422	14.8	245	5.3
Green areas	1,156	13.3	36	2.9	132	4.6	988	21.5
None	6,816	78.5	1,153	93.4	2,307	80.6	3,356	73.1

Table 3: Description of the neighborhood qualitative variables

	Total		Barcelona		Madrid		Valencia	
	N	%	N	%	N	%	N	%
Commerce								
Very good	671	7.7	34	2.8	118	4.2	519	11.3
Good	6,405	73.8	791	64.0	1,945	68.0	3,669	79.9
Regular	1,198	13.8	363	29.4	547	19.1	288	6.3
Bad	411	4.7	47	3.8	251	8.8	113	2.5
Neighborhood								
Very good	1,566	18.0	243	19.7	424	14.8	899	19.6
Good	6,130	70.6	957	77.5	1,805	63.1	3,368	73.4
Bad/Fair	989	11.4	35	2.8	632	22.1	322	7.0
Income								
High	644	7.6	25	2.0	344	12.0	295	6.4
Medium-high	1,884	21.7	192	15.6	531	18.6	1,161	25.3
Medium-Low	6,137	70.7	1,018	82.4	1,986	69.4	3,133	68.3

The hierarchical analysis is applied on four specifications; the first (baseline model) only includes an intercept. This specification will allow us to test if the three-level model is preferred to the single-level model; the second specification examine the influence of building characteristics (apartment and block) on price; the third model includes the neighborhood variables; while the fourth specification measures the influence of the city. In order to undertake the third-level-analysis, information must be available in the database about the apartments' postal code, as in our case, which can thus be used as a proxy for the neighborhood.

Our initial hypothesis is that there is greater homogeneity among apartments belonging to the same postal code and greater heterogeneity among apartments in different cities and/or costal codes.

5. EMPIRICAL ANALYSIS

This section discusses the results obtained from applying HLM and compares the results.

Our analysis considers four model specifications (Table 4). The specification of Model 1 only includes the intercept, with no explanatory variables for price. This is the base model and will be used to calculate the pseudo-R^2 (Snijders and Bosker, 1999; Giuliano *et al.*, 2010). Model 2 includes the building characteristics as explanatory variables, Model 3 incorporates neighborhood characteristics, and Model 4 use all building, neighborhood and city variables.

The comparison of the four HLM models is done by determining the significant variables and whether there are changes regarding the magnitude of the coefficients. Pseudo-R^2 is also compared, as is AIC and the reduction in total variance explained by inter-group differences, or the so called *variance partition coefficient*.

The results given in Table 4 show that all building variables are highly significant, regardless of the regression model estimated (with the exception of Number of apartments that becomes no significant when Neighborhoods variables are included). As pointed out in Tanaka *et al.*, (1982), the fact that the bedroom coefficient is negative is due to the strong correlation between this variable and Area. In the case of fixed floor space, the larger the number of rooms, the lower the price, since smaller rooms reduce prices. According with our estimations by Model 4, the elasticity price-area for an average apartment with three bedrooms is 0.80, while for an apartment with four bedrooms (average plus one standard deviation) that elasticity increases to 0.83. The more bedrooms are in the apartment the most valuable is an increment in its area. All other signs are as expected, the coefficients of Area and Floor are positive; the coefficient of Age of block has a negative sign, showing that the older the block, the lower the price; also qualitative variables Terrace and Lift have positive coefficients. “High” quality, “Very good” location and “Green areas and more” community spaces blocks are worth higher prices.

When considering neighborhood variables (Model 3), some variables included in Model 2 reduce their coefficient in absolute terms: the Number of apartments coefficient goes from 0.0099 to 0.0041 (becoming no significant); the coefficients of level Very good in Location goes from 0.237 to 0.174, and the coefficients of green areas and more in Community spaces goes from 0.09 to 0.07.

All neighborhood variables are significant and have the expected sign: neighborhoods with “Very good” commerce, “High” incomes or “Good” standards have higher priced apartments.

Model 4 includes Level 3 variable Mortgages, which is highly significant: an increment of 1% in the mortgages (as a proxy of the demand in the real state market) generates an increment of 0.9% in the prices.

As for the magnitude of the coefficients, no important differences are observed other than those already mentioned. The positive or negative sign is identical for all models.

The different models we developed make it possible to analyze the evolution of total variance and the variance partition coefficient. In this way we can calculate the percentage of the total variance explained by differences in cities (group-level 3) and postal codes (group-level 2), and by differences at the apartment level (individual-level).

We assume, and interested in testing, that the elasticity price-area is not constant between cities. More precisely, that the source of variation in apartment prices is different for each city. Thus, in models 2 to 4 we have included Area as a random coefficient to estimate how the between-apartments and within-apartments variance may change across the three cities. Therefore, the coefficient of Area or $\beta_k = \beta + v_{1k}$ has a constant slope which estimation is showed in Table 4 in row *In(Area)* and a random slope at level three which estimated variance is showed in Table 4 in section Random Effect as City variance (Area). Those terms and the estimation of the random slope are showed in Table 5. Regardless the model, the smallest elasticity price-area corresponds to Valencia and the largest to Barcelona. However, the difference in the elasticity between cities is small.

Table 4 also shows variance components (random effects): city variance, neighborhood variance and residual variance. The first is the level 3 variance generated by differences between cities not captured in the model. There is a constant random effect and a random slope of variable area effect of city. The second is the level 2 variance generated by differences between neighborhoods not captured in the model; and the third is the individual-level variance generated by differences between the apartments that are also not captured in the model. Thus, total city variance and therefore total variance of a model are not constant and depend on the area of the apartment. We have calculated both total variances for the average of the logarithm of the area and used these values to compute VPC. The results are showed in Table 4.

In Model 1 none explanatory variable is included, so that Total variance is the largest (and equal to 0.2825) and the residual variance (apartment-level variance) is responsible for as much as 47.1% of the total variance, followed by neighborhood variance (44.2% of the total variance). In Model 2 the building effects are included and Total variance drops drastically to 0.1266. The reason for this is that Model 2 included many important variables at the individual level. Moreover, the residual VPC drops to 29.1% and both neighborhood and city variances share the remaining variability. When neighborhood characteristics are included in Model 3, Total variance remains almost equal with respect to Model 2 but the neighborhood VPC drops by 10 points and almost 50% of the variability in the model corresponds to city variability. The inclusion in the model of the city variable mortgages has a severe impact on the Total variance that halves to 0.0645 and on city VPC that falls to 1.4%. Thus mortgages are able to explain almost all the variability at level 3. The remaining is share between residual variance (VPC equal to 53.2%) and neighborhood variance (VPC of 45.4%).

Table 4. Estimating the models and goodness of fit

	Model 1	Model 2	Model 3	Model 4
(Intercept)	12.57***	9.201***	9.135***	-1.997
	(0.0971)	(0.151)	(0.154)	(1.182)
ln(Area)		0.684***	0.689***	0.680***
		(0.0207)	(0.0238)	(0.0195)
ln(Floor)		0.0204***	0.0215***	0.0214***
		(0.00388)	(0.00375)	(0.00375)
ln(Bedrooms)		-0.599***	-0.533***	-0.545***
		(0.0679)	(0.0669)	(0.0652)
ln(Bathrooms)		0.157***	0.138***	0.138***
		(0.00876)	(0.00850)	(0.00850)
Terrace: Yes versus No		0.0817***	0.0839***	0.0843***
		(0.00752)	(0.00728)	(0.00728)
ln(Area)*ln(Bedrooms)		0.117***	0.105***	0.108***
		(0.0156)	(0.0154)	(0.0150)
ln(Number of apartments)		0.00990*	0.00411	0.00431
		(0.00411)	(0.00398)	(0.00398)
Age of block		-0.00233***	-0.00322***	-0.00321***
		(0.000252)	(0.000247)	(0.000247)
Lift: Yes versus No		0.117***	0.106***	0.106***
		(0.00623)	(0.00604)	(0.00604)
Location: Fair versus Bad		0.0777***	0.0751***	0.0731***
		(0.0146)	(0.0141)	(0.0141)
Location: Good versus Bad		0.143***	0.113***	0.110***
		(0.0165)	(0.0160)	(0.0160)
Location: Very good versus Bad		0.237***	0.174***	0.174***
		(0.0232)	(0.0226)	(0.0226)
Quality: Medium versus Low		0.104***	0.0842***	0.0840***
		(0.00572)	(0.00560)	(0.00560)
Quality: High versus Low		0.265***	0.201***	0.201***
		(0.0115)	(0.0115)	(0.0115)
Community spaces: Green areas vs. None		-0.00922	0.00925	0.00933
		(0.00703)	(0.00694)	(0.00693)
Community spaces: Green areas and more vs. None		0.0910***	0.0702***	0.0695***
		(0.0100)	(0.00979)	(0.00977)
Commerce: Regular versus Fair			-0.0219	-0.0216
			(0.0121)	(0.0121)
Commerce: Good versus Fair			-0.00205	-0.00262
			(0.0128)	(0.0128)
Commerce: Very good versus Fair			0.0510**	0.0506**
			(0.0157)	(0.0157)
Income: Medium-High vs. Medium-Low			0.0916***	0.0918***
			(0.00631)	(0.00631)
Income: High versus Medium-Low			0.154***	0.153***
			(0.0119)	(0.0119)
Neighborhood: Very Good versus Bad/Fair			0.0668***	0.0671***
			(0.00964)	(0.00964)
Neighborhood: Good versus Bad/Fair			0.0862***	0.0862***
			(0.0119)	(0.0119)
Mortgages				0.927***
				(0.0972)
Random effects:				

City variance (Const)	0.0245	0.0456	0.0491	0.0000
City variance (Area)	-	0.0000885	0.000536	0.0000459
Neighborhood variance	0.125	0.0424	0.0297	0.0293
Residual variance	0.133	0,0368	0.0343	0.0343
Total variance[1]	0.2825	0.1266	0.1239	0.0645
Level 3 VPC	8.7	37.4	48.3	1.4
Level 2 VPC	44.2	33.5	24.0	45.4
Level 1 VPC	47.1	29.1	27.7	53.2
Goodness of fit:				
Pseudo R^2		0.466	0.490	0.798
AIC	7,568.2	-3,531.8	-4,144.1	-4.155.7
Num. obs.	8,685	8,685	8,685	8,685
Num. groups: PC - City	106 - 3	106 - 3	106 - 3	106 - 3

*** $p< 0.001$, ** $p < 0.01$, * $p < 0.05$

[1] Total variance and VPC calculated in the average ln(Area).

Table 5. Estimated elasticities price-area.

Model	β	v_{1k}	Elasticity $\beta_k = \beta + v_{1k}$
Model 2	0.684	0.00739	0.691
		-0.00147	0.683
		-0.00592	0.678
Model 3	0.689	0.02718	0.716
		-0.00644	0.683
		-0.02075	0.668
Model 4	0.680	0.00609	0.686
		0.00036	0.680
		-0.00646	0.674

Table 6. Three model versus simpler models (LR test)

	LR	d.o.f	p-value
Three-level model versus single-level model (least square)	5355.79	2	0.0000
Three-level model versus two-level model (Postal code)	4001.63	1	0.0000
Three-level model versus two-level model (City)	9.60	1	0.0020

The log-likelihood ratio test indicates that the three-level-model increases significantly the likelihood with respect to the one-level-model (OLS), and with respect to both two-level-models, the one where PC is the second level and the one where the city is the second level (Table 6). In other words, the additional random-effects parameters included in the three-level-model with respect to more simple models are significant.

In light of these results, we can conclude that Model 4 is very accurate. We obtain a pseudo-R^2 of 0.798, a very satisfactory level if we compare it with that obtained in previous valuation studies (Fan *et al.*, 2006; Selim, 2009) or in our previous models. No further city related variables need be added to improve the valuation model because the improvement range is very low. Instead, efforts should be made to improve the description of the apartments and neighborhoods.

6. CONCLUSIONS

This paper applies a hierarchical linear model on residential real estate mass appraisal. The advantage of HLM compared to traditional hedonic regression models is that HLM explicitly considers that the price of the dwellings located in the same neighborhood is not independent. This is in accordance with reality, as all the apartments will be influenced by the characteristics of the neighborhood. Therefore, hedonic models obtain biased and inefficient results, as within-group correlations are not considered. Furthermore, HLM gives valuable information on the percentage of the variance error caused by each level in the hierarchical model.

In this research, three hierarchy levels were employed, being the first level the apartment and the block, the second the neighborhood and the third, the city. The models were estimated applying a vast database including 8.685 apartments located in the three major Spanish cities: Madrid, Barcelona and Valencia. 12 variables describe the flat and the block, whereas 3 variables describe the neighborhood and one variable describes the city. Both quantitative and qualitative variables are included.

Several models are applied and the results show an improvement of the goodness of fit of the model including all three hierarchic levels, with a pseudo R^2 of 0.798. An important conclusion of the study is that, in order to improve the model, more and more accurate variables should be used to describe the apartments and the neighborhoods. But it would not be efficient to include more information to describe the third level, the cities.

7. REFERENCES

Aitkin, M., Anderson, D., Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A, 144*, 148-161.

Antipov, E.A., Pokryshevsakaya, E.B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and CART-based approach for model diagnostics. *Expert Systems with Applications, 39,* 1772-1778. Doi:10.1016/j.eswa.2011.08.077.

Arribas, I., García, F., Guijarro, F., Oliver, J., Tamosiuniene, R. (2016). Mass appraisal of residential real estate using multilevel modelling. *International Journal of Strategic Property Management, 20 (1),* 77-87. Doi:10.3846/1648715X.2015.1134702

Aznar, J., Ferrís-Oñate, J., Guijarro, F. (2010). An ANP framework for property pricing combining quantitative and qualitative attributes. *Journal of the Operational Research Society*, *61 (5)*,740-755. Doi:10.1057/jors.2009.31

Basu, S., Thibodeau, T.G. (1998). Analysis of spatial autocorrelation in house prices. *Journal of Real Estate Finance and Economics, 17 (1)*, 61-85.

Brown, K.H., Uyar, B. (2004). A hierarchical linear model approach for assessing the effects of house and neighborhood characteristics on housing prices. *Journal of Real Estate Practice and Education, 7. (1)*, 15-23.

Cervelló, R., García, F., Guijarro, F. (2011).Ranking residential properties by a multicriteria single price model. J*ournal of the Operational Research Society*, *62 (11)*, 1941-1950.Doi:10.1057/jors.2010.170

Chasco, C., Le Gallo, J. (2013). The impact of objective and subjective measures of air quality and noise on house prices: A multilevel approach for downtown Madrid. *Economic Geography, 89 (2),* 127-148. Doi: 10.1111/j.1944-8287.2012.01172.x

D’Amato, M. (2007). Comparing rough set theory with multiple regression analysis as automated valuation methodologies. *International Real Estate Review, 10 (2),* 42-65.

Downes, T.A., Zabel, J.E. (2002). The impact of school characteristics on house prices: Chicago. *Journal of Urban Economics, 52,* 1-25.

Duncan, C., Jones, K., Moon, G. (1993). Do places matter? A multilevel analysis of regional variations in health–related behavior in Britain. *Social science and Medicine, 37*, 725-733.

Fagan, A. A., Wright, E. M., Pinchevsky, G. M. (2015). Exposure to violence, substance use, and neighborhood context. *Social science research, 49,* 314-326.

Fan, G.Z; Ong S.E., Koh, H.C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43 (12),* 2301-2315.

Farmer, M.C., Lipscomb, C.A. (2010). Using quantile regression in hedonic analysis to reveal submarket competition. *Journal of Real Estate Research, 32 (4)*, 435-460.

Feng, Q., Wu, G. L. (2015). Bubble or riddle? An asset-pricing approach evaluation on China's housing market. *Economic Modelling*, *46*, 376-383. Doi:10.1016/j.econmod.2015.02004

Ferreira, E., Sirmans, G. (1988). Ridge regression in real estate analysis. *The Appraisal Journal, 56 (3)*, 311-319.

García, N.; Gámes, M., Alfaro, E. (2008). ANN + GIS: An automated system for property valuation. *Neurocomputing, 71*, 733-742. Doi:10.1016/j.neucom.2007.07.031.

Gelman, A., Fagana, J., Kiss, A. (2007). An analysis of the New York City Police Department's “stop-and-frisk” policy in the context of claims of racial bias. *Journal of the American Statistical Association, 102 (479)*, 813-823.

Giuliano, G., Gordon, P., Pan, Q., Park, JY. (2010). Accessibility and residential land values: Some tests with new measures. *Urban Studies, 47 (14)*, 3103-3130. Doi:10.1177/0042098009359949.

Glaesener, M.L., Caruso, G. (2015) Neighborhood green and services diversity effects on land prices: Evidence from multilevel hedonic analysis in Luxembourg. *Landscape and Urban Planning, 143*, 100-111. Doi: 10.1016/j.landurbanplan.2015.06.008.

Gloudemans, R.J. (1999). Mass appraisal of real property. *International Association of Assessing Officers.*

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

Isakson, H.R. (2001). Using multiple regression in real estate appraisal. *The Appraisal Journal (3)*, 424-430.

Jones, K., Bullen, N. (1993). A multilevel analysis of the variation in domestic property prices: southern Englenad 1980-1987. *Urban Studies, 30 (8)*, 1409-1426.

Tsai, I-C. (2014). Ripple effect in house prices and trading volume in the UK housing market: New viewpoint and evidence. *Economic Modelling, 40*, 68-75. Doi:10.1016/j.econmod.2014.03.026

Kontrimas, V, Verikas, A. (2011). The mass appraisal of real estate by computational intelligence. *Applied Soft Computing, 11*, 443-448. Doi:10.1016/j.asoc.2009.12.2003.

Lee, C.C. (2009). Hierarchical linear modelling to explore the influence of satisfaction with public facilities on housing prices. *International Real Estate Review, 12 (3),* 252-272.

Lee, C.C., Lin, H.Y. (2014). The impacts of quality of the environment and neighbourhood affluence on housing prices: a three-level hierarchical linear model approach. *Asian Economic and Financial Review, 4 (5),* 588-606.

Leishman, C., Costello, G., Rowley, S., Watkins, C. (2013). The predictive performance of multilevel models of housing sub-markets: A comparative analysis. Urban Studies, 50 (6), 1201-1220. Doi: 10.1177/0042098012466603.

Narula, S.C., Wellington, J.F., Lewis, S.A. (2012). Valuating residential real estate using parametric programming. *European Journal of Operational Research, 217,* 120-128. Doi:10.1016/j.ejor.2011.08.014.

Orford, S. (2000). Modelling spatial structures in local housing dynamics: a multilevel perspective. *Urban Studies, 37*, 1643-1670.

Pestana Barros, C.,Gil-Alana, L.A., Chen, Z. (2014). The housing market in Beijing and delays in sales: A fractional polynomial survival model. *Economic Modelling, 42*, 296-300. Doi: 10.1016/j.econmod.2014.07.004.

Palmquist, R.B. (1984). Estimating the demand for the characteristics of housing. *Review of Economics and Statistics, 66*, 394-404.

Plakandaras, V., Gupta, R., Gogas, P., Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling, 45*, 259-267. Doi:10.1016/j.econmod.2014.10.050

R Core Team (2014). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria (http://www.R-project.org).

Raudenbush, S.W., Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59 (1),* 1-17.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*, 2843-2852. Doi:10.1016/j.eswa.2008.01.044.

Singh, J. (2014). Effect of school and home factors on learning outcomes at elementary school level: a hierarchical linear model. *Education 3-13, (ahead-of-print),* 1-24.

Snijders, T., Bosker, R. (1999). Multilevel analysis: an introduction to basic and advanced multilevel modelling. London: *Sage Publications*.

Tanaka, H., Uejima, S., Asai, K. (1982). Linear Regression Analysis with fuzzy model. *IEEE Transactions on Systems Man and Cybernetics, 12 (6),* 903-907.

Tay, D., Ho, D. (1991). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation & Investment, 10*, 252-541.

Tso, G. K., Guan, J. (2014). A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption. *Energy, 66,* 722-731.

Wang W., Rothschild, D., Goel, S., Gelman, A. (2014). Forecasting elections with non-representative polls. *International Journal of Forecasting.* In press.

Zeng, J.H., Peng, C-L., Chen M-C., Lee C-C. (2013). Wealth effects on the housing markets: Do market liquidity and market states matter? *Economic Modelling, 32*, 488-495. Doi: 10.1016/j.econmod.2013.02.021.