v. 3, n. 19, 2023

Scientific Journal of Applied Social and Clinical Science

MULTILEVEL METHODOLOGY APPLIED TO THE STUDY OF HOUSING SUPPLY IN BARCELONA: A COMPREHENSIVE ANALYSIS

Adrián González Jiménez
Universitat Politècnica de València
Spain
ORCID: 0009-0008-4388-6960

Alicia Llorca-Ponce
Universitat Politècnica de València
Spain
ORCID: 0000-0002-5426-4090

Soledad Valero Cubas
Universitat Politècnica de València
Spain
ORCID: 0000-0003-4565-326X

Abstract: El estudio del precio de una vivienda está asociado a variables que, si bien han sido utilizadas desde el s. XX como base de muchas metodologías, están sujetas a problemas dentro del análisis estadístico, principalmente la multicolinealidad. El objetivo de este trabajo es analizar la incidencia de diversas variables, internas y externas, en el precio de la vivienda con métodos de análisis estadístico que proporcionen información previa para descartar, entre otros, problemas de colinealidad.
Este análisis forma parte de una investigación más amplia que desarrolla una Red Neuronal Artificial (RNA) para valoración de viviendas en la ciudad de Barcelona. La selección de variables para esta fase del estudio fue extraída de una aplicación de reducción de datos (análisis factorial). Para afinar la adecuación de estas variables, se plantea la hipótesis de que otros métodos estadísticos verificarán la idoneidad de las variables previamente seleccionadas.
Verificamos la hipótesis a partir de un análisis multinivel aplicado a las variables estudiadas. El modelo estadístico pone de manifiesto la diferencia entre el efecto de las características individuales de las viviendas (primer nivel) y aquellas que provienen del entorno y que son comunes a todas las viviendas de un mismo barrio/distrito/ciudad (segundo nivel).
Keywords: Housing Prices, Real Estate Market, Multilevel Analysis, Barcelona

INTRODUCTION

The aim of this paper is to analyse a sample of 10 145 homes for sale in the city of Barcelona by means of multilevel models capable of providing information on the influence of different variables on price.

The usefulness of multilevel methods lies in their ability to analyse the influence of factors that are correlated at different scales. For the models in this study, descriptor variables for the properties (level 1) and descriptor variables for the neighbourhoods where they are located (level 2) have been used. These variables are related to each other, since the characteristics of the neighbourhood influence the individual value of each property. Although this assumption is contrary to the fundamentals required for linear regression analytical methods, multilevel models study the covariance between different variables to avoid problems arising from multicollinearity between variables.

The methodology applied in this study consists of a series of consecutive multilevel models, in a total of three stages, each time adding an additional variable to the model in an attempt to reduce the covariance estimated by the model, thus identifying variables that are largely responsible for the price variation between the dwellings in the sample. First, a null model is constructed from the dwellings separated by neighbourhoods; then a level 2 variable is added to the model and, finally, a further level 1 variable is added to the model.

Based on the models obtained, it can be seen that socio-economic variables such as the income index or the unemployment rate help to correct the covariance registered in the control models to a large extent and, to a lesser extent, the number of rooms in the dwelling.

LITERATURE REVIEW

The price of housing is conditioned by many factors, both intrinsic to the property and to the environment in which it is located. Since the mid-20th century, numerous studies have quantified the impact of external factors on house prices, such as proximity to green spaces (Tyrväinen, 1997), undesirable activities (Boyle & Kiel, 2001). Many of the models designed in this context to approximate house prices —extended from work on hedonic models (Rosen, 1974)— assume independence between variables.

However, hedonic models have problems of multicollinearity, where certain variables overlap in terms of their impact on the variance of the estimated results.

One family of statistical models that addresses this problem is the multilevel methodology (Goldstein, 2003), also known as random coefficient (Longford, 1993) or hierarchical (Raudenbush & Bryk, 2002) methods, which establish hierarchies among the variables. Hierarchical data structure is necessary when model variables are likely to be spatially correlated (Jones 1991). This is the case of the "neighbourhood effect", when we want to analyse the impact of several variables on house prices. Housing prices depend on house characteristics such as size, facilities or distribution, but it also depends on the characteristics of its surroundings (green areas, facilities...). In these cases, the price will subsume both characteristics of the property itself and shared features in the neighbourhood. Thus concluding that the aspects which add value to the property are shared with the other dwellings in the neighbourhood. The multilevel analysis will base on a hierarchy. At level 1, variables that are specific to each property, such as size or state of conservation; and at level 2, characteristics of the environment, shared by the houses in the neighbourhood.

The key aspect of multilevel models is the possibility for the parameters on which the dependent variable depends to vary. Jones and Bullen (1994) show that, for the analysis of prices in the City of London, multilevel models that recognise houses being located within boroughs with different characteristics are preferable to single-level models. Hierarchical models outperform traditional hedonic models, Kyung-Ku, Chun (2012). The hierarchical linear model is used to resolve autocorrelation and heteroscedasticity, Choi et al (2019).

For the present study, on housing prices in the city of Barcelona, level 1 will check number of rooms or distance to the city centre; and level 2, the neighbourhoods according to the Barcelona City Council delimitation. The value of level 2 variables, such as educational facilities or income level, is shared by all the dwellings in the neighbourhood. In the recent literature, we can find works in which multilevel models have been applied to study the influence of various factors on house prices.

Choi et al (2019) use a two-level hierarchical linear model with a multilevel model to analyse how various physical and environmental variables affect housing prices per square metre in the city of Busan, South Korea. Level 1 comprises physical characteristics of the house; at level 2, external variables, called walkability, measured for each basic district. Walkability is understood as a measure of the walkability of an area and valued by walking distance to facilities and amenities. The study concludes that walkability variables derive about 77% of the price.

Kyung-Ku, Chun (2012) focused on the influences of regional characteristics of urban infrastructure on housing prices, concluding that the hierarchical linear model can provide valuable information with important implications for urban policy such as urban infrastructure provision and balanced development.

Liu et al (2020) explore the housing price effects of ecological land, including forests, grasslands, wetlands and cultivated land, in Wuhan, China. The study shows that demand for forests, grasslands and wetlands can increase housing prices; cultivated land too, but to a lesser extent.

Hou (2016) use a multilevel hedonic model for single-family housing prices in the city of Los Angeles to determine whether traffic congestion negatively affects the price of single-family housing in the city of Los Angeles by worsening accessibility. The results suggest that housing demanders are willing to pay more for more accessible housing, as measured by commuting time, and that there are differences in the value of accessibility as a function of the income level of the population in the neighbourhoods. Tian et al (2017), who also study the impact of affordability on housing prices in Salt Lake County, United States, obtain different results. They consider that accessibility has two impacts: on the one hand, a positive one, that of accessibility to opportunities, but also a negative one, since the infrastructures and means of transport for such accessibility generate risks for environmental health, such as noise and air pollution. The results obtained indicate that the negative impacts (traffic noise and air pollution) are greater than the positive impact (accessibility). These results differ from those obtained in other studies applied to denser urban areas. Focusing on transport modes and specifically on rail transport, many and varied studies have tried to analyse the impact of a rail transit system on the value of housing in metropolitan areas. The studies, coinciding with the previous lines, have highlighted two effects, one positive, the improvements in accessibility due to reductions in travel times, and other negative effects due, especially, to those derived from noise and poorer views, Pan et al (2016). Studies have not agreed on the net effect of these two competing effects.

METHODOLOGY

The multilevel models studied here are based on a sample of homes for sale in the city of Barcelona taken from the Idealista real estate website. All the subjects used in this study have the same typology (flat in an apartment building); this building category accounts for 80.83% of the sample obtained, which comprises the 12550 dwellings for sale on the platform at the time the data were extracted.

Using the data obtained from the API of the real estate portal, the online services of the Barcelona City Council and some variables of our own elaboration, a database is created from which a total of 3 internal variables and 6 external variables are chosen by means of a factor analysis. The internal variables reflect intrinsic characteristics of each individual flat (crooms=rooms; exterior=orientation of the dwelling towards the interior/exterior of the block; SubwayDistance=linear distance from the dwelling to the nearest metro stop). The external variables define characteristics of the neighbourhood or district in which the properties are located: two of the variables collect data on existing facilities in the area (cPubSchools=public schools in the neighbourhood; cPrivSchools=private schools in the neighbourhood); the remaining 4 variables collect socio-economic indicators compiled by the Barcelona City Council (cIncomeIndex=income index of the neighbourhood compared to the rest of the city; cUnemployment=unemployment rate of the neighbourhood; cAreaAmenities=percentage of the commercial area of the neighbourhood devoted to *amenities*; cAreaIndustry=percentage of the commercial area of the neighbourhood devoted to industrial activity). All variables whose name begins with "c" have been centred.

From this selection of variables, a two-level hierarchical classification is made: level 1 includes internal variables, while level 2 includes external and socio-economic variables. For all models, housing price is used as the dependent variable. Once the classification has been established, several multilevel models are constructed to contrast the results with each other and to analyse the oscillations in the variance of house prices as a function of the factors studied in each model.

The programme used to obtain the models is SPSS ver. 16.

As a reference, we first obtain a null model (AEA = ANOVA with random effects factor) to analyse the variability of house prices. The neighbourhoods of Barcelona (73) are taken as a factor to group the dwellings in the sample. The variance of the "neighbourhood" factor is studied with respect to the variance of the residuals and is used as a test of the null hypothesis.

The results of these models are used to test the intra-class correlation coefficient (ICC): this parameter is obtained as the percentile of the variance of the "neighbourhood" factor with respect to the total variance included in the model. The lower the ICC, the lower the similarity between dwellings in the same neighbourhood.

From these models, the impact of level 1 variables is studied in those models where the ICC has given lower results, generating another 6 derived models (ACEA = ANCOVA with random effects factor): three models for each level 2 variable screened, depending on whether the number of rooms, the orientation of the dwelling or the distance to the metro is chosen as a covariate.

RESULTS

The distribution of dwellings by neighbourhood is fairly homogeneous despite the observed standard deviation, with the exception of four city neighbourhoods for which less than 10 cases are available. Descriptive statistics are attached at the end of the article in an appendix to check the distribution of the sample.

At a descriptive level (Table 1), the sample's average housing price is €521 331 with a standard deviation of around €480 000 between neighbourhoods, showing an apparent relationship between price and neighbourhood. This shows a segregation of high and low incomes in different sectors of the city.

(SD = standard deviation; CV = coefficient of variation)				
Neighbourhoods	Count	Media	DT	CV
73	10 145	521 331	480 513	92,2%

Table 1. Summary of descriptive statistics for the dependent variable *price* in each neighbourhood

The results of the AEA model (Table 2) allow us to reject the hypothesis that the "neighbourhood" factor has a null effect (Sig. < 0,000). Based on the estimation results, we can say that the "neighbourhood" factor accounts for 36% of the total price variability. This result is congruent with the standard deviation observed in the descriptive statistics of the sample used: in many of the city's neighbourhoods it is possible to find homes whose price varies greatly with respect to the average price of the area, both the minimum and the maximum.

The RMR models introduce the level 2 variables in the estimation. Based on these models, and for each of the fixed effects estimates they represent (Table 3 and 3b), it is observed that the average house price:

- decreases by €8953 for each additional public school in the neighbourhood;
- increases by €14 955 for each additional public school;
- increases by €6374 for each additional point in the neighbourhood income index;
- decreases by €79 638 for each additional percentage point in the unemployment rate in the neighbourhood;
- increases by €6499 for each additional percentage point of retail space in the neighbourhood allocated to *amenities*;
- and decreases by €8186 for each additional

Parameter	Estimate	Standard error	gl	t	Sig.	95% confidence interval	
						Lower limit	Upper limit
Intersection	4,063055E5	3,429480E4	67,436	11,847	,000	3,378610E5	4,747501E5

Table 2. Summary of the AEA model. Fixed effects estimates (dependent variable: *price*)

Model	Parameter	Estimate	Standard error	gl	t	Sig.
AEA	intersection	406 306	34 295	67,436	11,847	,000
RMR			**cPubSchools**			
	intersection	446 692	31 625,1	65,516	14,125	,000
	cPubSchools	-8953,0	2052,6	64,361	-4,362	,000
			cPrivSchools			
	intersection	444 347	33 262,2	65,688	13,359	,000
	cPrivSchools	14 955	4221,1	65,188	3,543	,001
			cIncomeIndex			
	intersection	495 025	11 103,8	63,381	44,582	,000
	cIncomeIndex	6374,0	253,2	64,538	25,170	,000
			cUnemployment			
	intersection	450 877	26 796,1	63,647	16,286	,000
	cUnemployment	-79 638	11 418,1	74,013	-6.975	,000
			cAreaAmenities			
	intersection	427 357	36 741,3	67,152	11,632	,000
	cAreaAmenities	6499,0	4335,9	71,112	1,499	,138
			cAreaIndustry			
	intersection	430 270	32 673,5	65,516	13,169	,000
	cAreaIndustry	-8186,0	2393	76,137	-3,421	,001

Table 3. Fixed effects of RMR models

Model	Parameter		Estimate	Typical error	Wald Z	Sig.	CCI	CV
RMR	cPubSchools						0,30	22,6
var level 2		residual	13,6E10	1,9E9	70,992	,000		
cPubSchools	neighbourhood	variance	5,9E10	10,6E9	5,506	,000		
	cPrivSchools						0,32	15,9
var level 2		residual	13,6E10	1,9E9	70,992	,000		
cPrivSchools	neighbourhood	variance	6,3E10	11,5E9	5,525	,000		
	cIncomeIndex						0,04	**92,2**
var level 2		residual	13,6E10	1,9E9	71,017	,000		
cIncomeIndex	neighbourhood	variance	6E9	1,3E9	4,686	,000		
	cUnemployment						0,24	**43,4**

var level 2		residual	13,6E10	1,9E9	70,986	,000		
cUnemployment	neighbourhood	variance	4,3E10	8E9	5,368	,000		
	cAreaAmenities						0,35	1,9
var level 2		residual	13,6E10	1,9E9	70,992	,000		
cAreaAmenities	neighbourhood	variance	7,4E10	1,3E10	5,562	,000		
	cAreaIndustry						0,32	13,6
var level 2		residual	13,6E10	1,9E9	70,990	,000		
cAreaIndustry	neighbourhood	variance	6,5E10	1,2E10	5,512	,000		

Table 3b. Covariance parameters (RMR) for each model according to the level 2 variable (var level 2) chosen.

Model	Parameter	Estimate	Standard error	gl	t	Sig.
ACEA						
crooms	**cIncomeIndex**					
	Intersection	486 538	11 349,9	64,692	42,867	,000
	cIncomeIndex	5733,8	259,1	66,093	22,127	,000
	crooms	95 537,9	2508,2	10 140,152	38,091	,000
	cUnemployment					
	Intersection	447 161	24 629,3	64,054	18,156	,000
	cUnemployment	-72 265,6	10 509,5	74,398	-6,876	,000
	crooms	96 144,2	2512,8	10 103,212	38,261	,000
outside	**cIncomeIndex**					
	Intersection	484 373	13 155,4	124,516	36,819	,000
	cIncomeIndex	6373,1	253,1	64,615	25,181	,000
	outside	13 466,8	8927,5	10 125,470	1,508	,131
	cUnemployment					
	Intersection	440 263	27 685,6	72,813	15,902	,000
	cUnemployment	-79 726,1	11 408,1	74,023	-6,989	,000
	outside	13 455,1	8940,1	10 088,807	1,505	,132
cSubwayDistance	**cIncomeIndex**					
	Intersection	493 921	11 190,5	63,087	44,137	,000
	cIncomeIndex	6350,9	255,2	64,232	24,891	,000
	cSubwayDistance	82 903,9	37 141,7	9580,691	2,232	,026
	cUnemployment					
	Intersection	449 302	26 767,7	63,450	16,785	,000
	cUnemployment	-79 302	11 404,6	73,742	-6,954	,000
	cSubwayDistance	96 026,3	37 616,8	10 134,538	2,553	,011

Table 4. ACEA models, fixed effects (level 1 variables in first column, level 2 in second column)

Model	Parameter		Estimate	Typical error	Wald Z	Sig.	CORR
ACEA							
crooms	**cIncomeIndex**						-7,939
		residual	1,2E11	1,7E9	71,013	,000	12,609
	Intersection [subject=neighbourhood]	variance	6,4E9	1,3E9	4,863	,000	
	cUnemployment						15,669
		residual	1,2E11	1,7E9	70,984	,000	12,583
	Intersection [subject=neighbourhood]	variance	3,6E10	6,7E9	5,380	,000	
outside	**cIncomeIndex**						0,131
		residual	1,4E11	1,9E9	71,014	,000	0,033
	Intersection [subject=neighbourhood]	variance	5,9E9	1,3E9	4,688	,000	
	cUnemployment						0,2
		residual	1,4E11	1,9E9	70,983	,000	0,009
	Intersection [subject=neighbourhood]	variance	4,3E10	8E9	5,369	,000	
cSubwayDistance	**cIncomeIndex**						-1,684
		residual	1,4E11	1,9E9	71,012	,000	0,069
	Intersection [subject=neighbourhood]	variance	6,1E9	1,3E9	4,682	,000	
	cUnemployment						0,277
		residual	1,4E11	1,9E9	70,982	,000	0,050
	Intersection [subject=neighbourhood]	variance	4,3E10	8E9	5,358	,000	

Table 4b. Covariance parameters (ACEA)

percentage point of commercial space in the neighbourhood devoted to industrial activity.

None of the models show significant variation in the residual values: variability is not affected by the introduction of level 2 (neighbourhood) covariates.

In the case of the variability associated with the neighbourhood, however, the impact of the covariates is in some cases absolute: when introducing the variables cIncomeIndex or cUnemployment into the model, the CCI drops from 36% to 4% and 24% respectively. It is concluded that the differences in average house prices for each neighbourhood are 92% attributable to the neighbourhood's income index and 43% to its unemployment rate.

Finally, the ACEA models (Table 4 and 4b) extend the results of the previous step by incorporating the internal variables crooms, exterior and cSubwayDistance. In terms of fixed effects, the internal variables always correct the estimation positively. In terms of the covariance parameters, they barely correct the initial variance. Adding the variables crooms and cSubwayDistance to the variable cIncomeIndex increases the variance by 8 and 1,7% respectively; all other combinations reduce the variance by less than 1%.

The effect of adding a level 1 variable to the multilevel model is, however, best seen in the model that studies the level 2 variable cUnemployment accompanied by the level 1 variable crooms. In other words, by combining the unemployment rate of a neighbourhood with the number of rooms in each dwelling, the variability of the average price between neighbourhoods has been corrected by 15,7%. Likewise, the within-neighbourhood price variability (the residual of the null model) has been corrected by 12,6% by incorporating crooms into the model with respect to the null model. In contrast to the rest of the models, this variability has also been corrected by 12,6% in the model with the combination cIncomeIndex + crooms. In the other four models, less than 1% is corrected (Table 4b, column "CORR").

DISCUSSION

These results aid to study the influence of different variables on the average price of the sample of dwellings used. In general terms, the null model has corroborated that there are variations in the average price of the different neighbourhoods of the city. Subsequently, it has been shown that this variability can be attributed, at least in part, to descriptor variables of the neighbourhoods and, ultimately, to descriptor variables of each individual dwelling.

The models obtained, have not responded in their entirety as initially proposed. The starting hypothesis was that the variability of the average price of the neighbourhoods could be corrected gradually as variables were identified to which part of this variance could be attributed: in the null model, the highest variance attributable to the neighbourhood factor would be obtained and, in more complex models carried out subsequently, each variable added would correct a higher percentage.

This has only been the case for the influence of the unemployment rate of the neighbourhoods when corrected for the number of rooms in the dwellings: the variance of the (neighbourhood) intercept decreased by 43% when corrected for unemployment, and then by a further 15,7% when corrected for the number of rooms. We can thus say that both variables have a significant influence on house prices and can explain a considerable percentage of the price variability within neighbourhoods.

In the case of the other models, however, the effect obtained is less evident. The reason may be that the different potential combinations of variables explain different sectors of variability which, between different models, may overlap. The variance of the covariance parameters in the four models with the lowest correction is very low, but comparable across them all: the different combinations of unemployment/income index with orientation/distance to the metro explain equivalent percentages of the variance of the model and are not sufficient to attribute a differentiated impact to each of them.

As a result, a future line of work has been opened up to explore more complex multilevel models in order to better understand the interactions of the variables studied.

On the one hand, it is possible to build alternative RMR and ACEA models with a larger number of variables in each of the iterations.

On the other hand, there are multilevel models (RCA = random coefficient regression analysis, RMPR = regression analysis of means and slopes as outcomes), which study the effect of the variables in more detail. RCA models take the linear regressions of the models studied in this paper and transform the intercepts and slopes of the regressions for each neighbourhood so that they can vary randomly (where previously these were zero or assumed to be fixed). The RMPR models, on the other hand, allow us to study the variation detected in the RCA models but attributing it to variables at both levels.

REFERENCIAS

Boyle, M. & Kiel, K.A. (2001): **"A Survey of House Price Hedonic Studies of the Impact of Environmental Externalities."** *Journal of Real Estate Literature* vol. 9, no. 2.

Choi, Y., Seo , M., & Oh, S. (2019). **The Correlates between Walkable Environments and Housing Price using Multi-level Model.** *KSCE Journal of Civil Engineering, 23*(3), 4516-4524. doi:10.1007/s12205-019-1894-0.

Goldstein, H. (2003). **Multilevel statistical models** (3rd ed.). *New York: Halstead Press*.

Hou, Y. (2016). **Traffic congestion, accessibility to employment, and housing prices: A study of single-family housing market in Los Angeles County.** *Urban Studies, 54*(15), 3423-3445. doi:10.1177/0042098016675093.

Jones, K. (1991). **Specifying and estimating multilevel models for geographical research.** *Transactions of the Institutte of British Geographers, 16*(2), 148-159. doi:https://www.jstor.org/stable/622610?origin=crossref

Jones, K., & Bullen, N. (1994). **Contextual Models of Urban House Prices: A Comparison of Fixed- and Random-Coefficient Models Developed by Expansion.** *Economic Geography, 70*, 252-272. doi:10.2307/143993.

Kyung-Ku, C. (n.d.). **A Study of the Influences of Urban Infrastructure on Housing Prices : A Hierarchical Linear Model.** *Journal of Korea Planning Association, 47*(4), 193-204. doi:https://www.webofscience.com/wos/alldb/full-record/KJD:ART001691693

Liu, T., Hu, W., Song, Y., & Zhang, A. (2020). **Exploring spillover effects of ecological lands: A spatial multilevel hedonic price model of the housing market in Wuhan, China.** *Ecological Economics, 170*. doi:https://doi.org/10.1016/j.ecolecon.2019.106568

Longford, N.T. (1993). **Random coefficient models.** *New York: Oxford University Press*.

Pan, Q., Pan, H., & Zhang, M. (2014). **Effects of rail transit on residential property values : Comparison study on the rail transit lines in Houston, Texas, and Shanghai, China.** *Transportation Research Record: Journal of the Transportation Research Board, 2453*(1), 118-127. doi:https://doi.org/10.3141%2F2453-15

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., and Congdon, R. (2000). **HLM 5: Hierarchical linear and nonlinear modelling.** Lincolnwood, IL: *Scientific Software International.*

Rosen, S. (1974): **"Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"**, *Journal of Political Economy vol. 82*, no. 1, pp. 34-55.

Tian, G., Wei, Y., & Li, H. (2017**). Effects of accessibility and environmental health risk on housing prices: a case of Salt Lake County, Utah.** *Applied Geography, 89*, 12-21. doi:10.1016/j.apgeog.2017.09.010.

Tyrväinen, L. (1997): **"The Amenity of the Urban Forest: An Application of the Hedonic Pricing Method."** *Landscape and Urban Planning*, vol. 37, no. 3-4, pp. 211-222.