Economics Bulletin

Volume 32, Issue 2

Hierarchy and spatial autocorrelation effects in hedonic models.

Coro Chasco
Universidad Autónoma de Madrid

Julie Le Gallo
Université de Franche-Comté

Abstract

Hedonic housing price models should deal with spatial autocorrelation in order to avoid bias and inconsistency in the coefficient estimates. Multilevel models have been presented as a way of properly considering the effects of neighbourhood amenities operating at different spatial hierarchies. In this paper, we test this feature by specifying a three-level model for a dataset on downtown Madrid, which includes spatially lagged explanatory variables. The results show that hierarchical models are not capable of fully capturing spatial autocorrelation so that more effort should be done in developing the appropriate spatial multilevel models.

This work has been carried out with the financial support of the Spanish Ministry of Education and Science ECO2009-10534. Coro Chasco also acknowledges financial support from Xunta de Galicia 2009/AX702.
Citation: Coro Chasco and Julie Le Gallo, (2012) "Hierarchy and spatial autocorrelation effects in hedonic models.", *Economics Bulletin*, Vol. 32 No. 2 pp. 1474-1480.
Contact: Coro Chasco - coro.chasco@uam.es, Julie Le Gallo - jlegallo@univ-fcomte.fr.
Submitted: October 29, 2011. **Published:** May 15, 2012.

1. Introduction

The hedonic price model is one of the most widely used methods for evaluating the marginal willingness to pay of households for various residences characteristics. The method consists in including in the regression model, as explanatory variables, the full set of significant determinants of housing prices, such as structural attributes of the properties, the characteristics of the immediately surrounding social and natural environment and accessibility or locational characteristics (Bowen et al. 2001). In this paper, we focus on two methodological issues, that have been analyzed separately in the literature.

On the one hand, the spatial arrangement of observations is largely of concern as the presence of omitted spatial dependence implies inconsistent and/or biased coefficient estimates and unreliable statistical inference (Wilhelmsson 2002). There are two main specifications that aim at modelling spatial autocorrelation. They are based on the use of a spatial weights matrix, which expresses the form of spatial connectivity among each pair of observations. The first model includes spatial lags of the dependent – and sometimes also explanatory variables –, whereas in the second spatial autocorrelation is directely modelled in the error term (Anselin 1988; LeSage and Pace 2009; Le Gallo 2013).

On the other hand, multilevel models allow considering the effects of neighborhood amenities operating at different hierarchies or upper-scaled spatial levels. They are capable of 'contextualizing' the hedonic specification by allowing housing implicit prices to vary between different clusters or submarkets.

In this context, the aim of this paper is to analyze whether spatial autocorrelation in hedonic prices is properly captured by specifying a multilevel model. This issue is of some importance as the literature in spatial econometrics has pointed out that spatial heterogeneity and spatial autocorrelation entertain complex links (Anselin and Bera 1998). In particular, Brunsdon et al. (1999) argue that spatial autocorrelation is sometimes the result of unmodelled parameter instability. In other words, if space-varying relationships are modelled within a global regression, the error terms may be spatially autocorrelated. In the context of multilevel models, spatial heterogeneity is modelled through random coefficients operating at the different geographical levels. Following this idea, Orford (2000) indeed argues that multilevel models properly capture spatial effects since they control for spatial heterogeneity. However, Morenoff (2003), who was interested in testing and interpreting spatial dependence in health, found that some spatially-lagged explanatory variables were still significant in a two-level hierarchical model.

To investigate the ability of multilevel models in fully capturing spatial autocorrelation effects in the errors, we first specify a three-level hierarchical model, which is more complete and capable of explaining data heterogeneity present in three different spatial scales. Additionally, we add the set of spatially lagged regressors, as pure control variables in order to show that there is still a need to take care of spatial autocorrelation in multilevel models, whereas Morenoff is only interested in their parameter interpretation. This model is applied on a dataset on downtown Madrid. The analysis of the significance of these variables indicates whether specifying a multilevel model is enough to capture all the spatial autocorrelation present in the sample. The next section presents the data and the results while the last section concludes.

2. Application to downtown Madrid

Our study focuses on the city center of Madrid, 'Central Almond', which is an area administratively formed by 43 neighbourhoods and 780 census tracts, encircled by the first metropolitan ring-road (the M30). Due to confidentiality constraints, it is difficult to obtain housing prices microdata from Spanish official institutions. For this reason, our records were drawn from an on-line real estate database, 'idealista.com', and refer to January 2008. The asking price has then been used as a proxy for the selling price. In total, around 5,080 housing prices were finally recorded after the corresponding consolidation and geocoding processes.

At the first level of houses, we employ ten explanatory variables, seven of which are attribute data provided by 'idealista.com'; i.e. floor level, dwelling type, living space, and modernization and repair (see Table I for a complete definition). Additionally we constructed two accessibility measures (distance to the financial district and distance to the main road-axis and commercial avenues) since they are frequently advertised by real estate agents and often capitalized in housing prices. As air-pollution is an important determinant of house prices (Smith and Huang 1995), another explanatory variable in our model is the percentage of households that estimate that their homes' surroundings are polluted. Since it is a 2001 Census variable, which is only available at the level of census tracts, it has been interpolated at the level of houses.

Finally, our model also includes other variables at a second level of census tracts. There are three Census variables on socioeconomic and demographic characteristics related to home-ownership: percent of people aged over 65, percent of population with secondary and university degrees and percent of unemployed people.

Table I. The variables used in the model

Variable	Description	Source	Units	Period
LEVEL 1: HOUSES				
lprice	Housing price	Idealista	Euros (in logs)	Jan. 2008
floor	First floor and upper	Idealista	0-1	Jan. 2008
attic	Attic	Idealista	0-1	Jan. 2008
house	House ('chalet')	Idealista	0-1	Jan. 2008
duplex	Duplex	Idealista	0-1	Jan. 2008
bedsit	Bedsit	Idealista	0-1	Jan. 2008
reform	Old house that must be reformed	Idealista	0-1	Jan. 2008
lm2	Living space	Idealista	Square meter (in logs)	Jan. 2008
axis	Proximity to the main axis	Self-elab.	0-1	-
discen	Distance to the financial district	Self-elab	Km.	-
cont	Subjective air-pollution indicator	Census	%	Nov. 2001
LEVEL 2: CENSUS TRACTS				
p65	Percent of population over 65 years	Padrón, INE	%	Jan. 2008
educ	Education level (secondary/university)	Census, INE	-	2001
unem	Unemployment rate	Census, INE	-	2001

We first specify a three-level spatial model which allows determining how variations in housing prices are allocated across each spatial level: houses, census tracts and neighbourhoods. The model is in semi-log form:

$$\begin{cases} lprice_{ijk} = \beta_{0,jk} + \beta_{1,j} lm2_{ijk} + \beta_{2,k} cont_{ijk} + \sum_{s=1}^{S} \alpha_s x_{s,ijk} + \varepsilon_{ijk} \\ \beta_{0,jk} = \gamma_{00,k} + \sum_{l=1}^{N_0} \gamma_{0l} x_{0l,jk} + w_{0,jk} \\ \gamma_{00,k} = \mu_{000} + u_{00,k} \\ \beta_{1,j} = \gamma_{10} + w_{1,jk} \\ \beta_{2,k} = \mu_{200} + u_{20,k} \end{cases} \quad (1)$$

where $lprice_{ijk}$ is the log of price of transaction *i* in census tract *j* and neighborhood *k*; *lm2* and *cont* are floor area and air-pollution variables and *S* is the number of the rest of property-level structural and accessibility attributes, $x_{s,ijk}$, whose associated coefficients (α_s) are assumed to be fixed across upper-level spatial scales. This is not the case of floor area and air-pollution estimates ($\beta_{1,j}$, $\beta_{2,k}$), which are allowed to vary randomly across their corresponding census tract and neighbourhood means (γ_{10}, μ_{200}), respectively. A significant spatial variation in the implicit prices of these variables would indicate the existence of different sub-markets inside Central Almond.

The intercept of the model in level 1 ($\beta_{0,jk}$) is affected by a group of N_0 attribute variables at the census tract level ($x_{0l,jk}$). They are *p*65, *educ* and *unem* (see Table I), whose effects on housing prices (γ_{0l}) are assumed to be fixed across census tracts, i.e. they do not vary randomly at the neighborhood level. $\gamma_{00,k}$ is the neighborgood *k*'s specific constant μ_{000} is the general intercept of the model.

Regarding the error terms, $\varepsilon_{ijk} \sim Nid\left(0,\sigma_{\varepsilon}^2\right)$ is the random term measuring the deviation of transaction *ijk*'s log of price from the mean log of price in census tract *j*; $w_{0,jk} \sim Nid\left(0,\sigma_{w0}^2\right)$ is the random term measuring the deviation of census tract *jk*'s mean log of price from the mean log of price in neighborhood *k*; and $u_{00,k} \sim Nid\left(0,\sigma_{u0}^2\right)$ is the random term measuring the deviation of neighborhood *k*'s mean log of price from the grand mean.

As reported in Table II, all the estimates are strongly significant and show the expected sign. The likelihood-ratio (LR) test of absence of random effects strongly rejects the null, hence suggesting that a multilevel approach with random effects is indeed relevant. We found significant variations of housing prices at the level of the individual transactions (σ_{ε}^2), census tracts (σ_{w0}^2) and neighborhoods (σ_{u0}^2). This model further enables exploring the importance of floor area (*lm2*) –the most important property attribute– and air-pollution (*cont*) in house price variation by allowing their marginal prices (parameters) to vary at the neighbourhood level, in the case of air-pollution (σ_{u20}^2) or census tract level, for floor area (σ_{w1}^2). The covariance terms between the constant and floor area/air-pollution ($\sigma_{w0,w1}$, $\sigma_{u00,u20}$) are significant and negative, as expected; they measure the relationship between the price of floor area/air-pollution and average census tract/neighbourhood-level house price, respectively.

Table II. Multilevel model REML estimation results

			Non-spatial	Spatial
	Constant		8.948862**	8.295584**
Structural	*floor*		0.120377**	0.119603**
	attic		0.052906**	0.053721**
	house		0.262311**	0.273191**
	duplex		0.050497**	0.054259**
	bedsit		0.060970**	0.060814**
	lm2		0.871471**	0.867382**
	reform		-0.097873**	-0.097319**
Accesibility	*axis*		0.036292**	0.028422**
	discen		-0.050406**	-0.049466**
Pollution	*cont*		-0.004407**	-0.003018**
Census tracts	*p65*		-0.004474**	-0.002673*
	educ		0.005733**	0.003139**
	unem		-0.005415**	-0.002694
Spatial lags	*Wfloor*		-	-0.013493
	Wattic		-	-0.061602
	Whouse		-	-0.096131
	Wduplex		-	0.137592
	Wbedsit		-	0.396004**
	Wlm2		-	0.161565**
	Wreform		-	0.114322
	Wp65		-	-0.006379**
	Weduc		-	0.003570**
	Wunem		-	-0.008829*
Variance and covariance (standard error)	*Neighb.*	*constant* (σ^2_{u00})	0.026928 (0.01127)	0.023610 (0.00965)
		cont (σ^2_{u20})	0.000013 (8.26e-06)	0.000014 (7.33e-06)
		cont vs. constant ($\sigma_{u00,u20}$)	-0.000573 (0.00031)	-0.000573 (0.00027)
	Census	*constant* (σ^2_{w0})	0.203499 (0.02907)	0.201130 (0.02854)
		lm2 (σ^2_{w1})	0.064993 (0.00846)	0.062612 (0.00825)
		lm2 vs. constant ($\sigma_{w0,w1}$)	-0.114270 (0.01562)	-0.111606 (0.01529)
	Houses	(σ^2_{ε})	0.024690 (0.00055)	0.024712 (0.00055)
Log-likelihood:			1,689.07**	1702.90**
Deviance (H_0: Non-spatial model)			-	27.65**

In brackets are the standard errors, * significant at 0.05, ** significant at 0.01

The analysis up to this point has taken into account the contextual effects influencing housing prices through the consideration of attributes at upper-level spatial scales as well as the variation of implicit prices of some goods across census tracts and neighbourhoods in downtown Madrid. If spatial autocorrelation was properly taken into account, after controlling for these effects in a multilevel regression model with hierarchical random effects, additional spatially lagged variables should not be significant. In order to test this hypothesis, we specify a cross-regressive multilevel model, introducing in Eq. (1) the spatial lag terms for all the individual and contextual explanatory variables[1] as follows:

$$\begin{cases} lprice_{ijk} = \beta_{0,jk} + \beta_{1,j} lm2_{ijk} + \beta_{2,k} cont_{ijk} + \sum_{s=1}^{S} \alpha_s x_{s,ijk} + \lambda_1 Wlm1_{ijk} + \sum_{s=1}^{S-2} \delta_s Wx_{s,ijk} + \varepsilon_{ijk} \\ \beta_{0,jk} = \gamma_{00,k} + \sum_{l=1}^{N_0} \gamma_{0l} x_{0l,jk} + \sum_{l=1}^{N_0} \pi_{0l} Wx_{0l,jk} + w_{0,jk} \\ \beta_{1,j} = \gamma_{10} + w_{1,jk} \\ \gamma_{00,k} = \mu_{000} + u_{00,k} \\ \beta_{2,k} = \mu_{200} + u_{20,k} \end{cases} \quad (2)$$

To construct the spatial lags, we use an inverse squared distance matrix. The difference in the likelihood ratio statistic of this model and the previous one (i.e. the deviance or likelihood ratio test) is 27.65. Under the null hypothesis, it follows a chi-squared distribution with degrees of freedom equal to 10, i.e. the number of new parameters (Woodhouse et al, 1996). The *p*-value is less than 0.002: the spatial lags of the explanatory variables therefore have globally a significant effect in explaining house price variation in the model. Looking at individual variables, it appears that 5 of these spatially lagged explanatory variables are significant at 5% (4 at 1%). The interpretation of the associated coefficients is not sraightforward and is out of the scope of that paper. However, their significance prove the existence of remaining spatial autocorrelation in the regression.

3. Conclusion

In this paper we highlight the need of introducing spatial autocorrelation processes in multilevel models. The results prove that spatial multilevel models are not capable of fully capturing all the spatial processes present in a social variable, such as housing prices. Consequently, more effort should be done in order to develop appropriate spatial multilevel models in order to deal with spatial autocorrelation. Elhorst and Zeilstra (2007) advance a first solution for a two-level model, correcting for spatial error dependence among observations (regions, in their case) within different groups (countries). Nevertheless, further research should be done in order to estimate multilevel spatial lag models as well as to develop adequate Lagrange Multiplier missepecification tests on spatial autocorrelation in the error terms; in cross-section spatial models, these tests are the basis of the so-called classical approach to econometric modeling (Florax et al 2003), allowing for the selection of a correct spatial specification between multilevel spatial lag models and multilevel spatial error models.

References

Anselin, L. (1988) *Spatial econometrics: Methods and models*, Kluwer Academic: Dordrecht.

Anselin, L. and A.K. Bera (1998) "Spatial dependence in linear regression models with an application to spatial econometrics" in *Handbook of Applied Economics Statistics* by Ullah A. and D.E.A. Giles, Eds., Springer: Berlin, 21-74

[1] Since air-pollution is a spatially interpolated variable and the accessibility indicators are built as distances to fixed points, their corresponding spatial lags produce extreme multicollinearity in the spatial cross-regressive model. For this reason, these spatial lags have been omitted from this specification.

Bowen, W.M., B.A. Mikelbank and D.M. Prestegaard (2001) "Theoretical and empirical considerations regarding space in hedonic housing price model applications", *Growth and Change* **32**, 466-490.

Brunsdon, C., A.S. Fotheringham and M. Charlton (1999) "Some notes on parametric significance tests for geographically weighted regression" *Journal of Regional Science* **39**, 497-524.

Elhorst, J.P. and A.S. Zeilstra (2007) "Labour force participation rates at the regional and national levels of the European Union: An integrated analysis" *Papers in Regional Science* **86**, 525-550

Florax, R.J.G.M., H. Folmer and S.J. Rey (2003) "Specification searches in spatial econometrics: the relevance of Hendry's methodology" *Regional Science and Urban Economics* **33**, 557–579.

Le Gallo, J. (2013) "Cross-section spatial regression models" in *Handbook of Regional Science* by Fischer M.M. and P. Nijkamp, Eds., Springer, Berlin, in press.

LeSage, J. and R.K. Pace (2009) *Introduction to Spatial Econometrics*, CRC Press.

Morenoff, J.D. (2003) "Neighborhood mechanisms and the spatial dynamics of birth weight", *American Journal of Sociology* **108**, 976-1017.

Orford, S. (2000) "Modelling spatial structures in local housing market dynamics: a multilevel perspective", *Urban Studies* **37**, 1643-1671.

Smith, V.K. and J.C. Huang (1995) "Can markets value air quality? A meta-analysis of hedonic property value models", *Journal of Political Economy* **103**, 209-227

Wilhelmsson, M. (2002) "Spatial models in Real Estate economics", *Housing, Theory and Society* **19**, 92–101

Woodhouse, G., Rasbash, J., Goldstein, H. and M. Yang (1996) "Introduction to multi-level modelling" in *Multi-level Modelling Applications: A Guide for Users of Mln* by Woodhouse, G., Ed., Institute of Education, University of London, 9–57