Measuring the Value of a Moscow Apartment: a Spatial Approach to the Hedonic Pricing of Attributes

Repkine, Alexandre

Konkuk University

March 2008

Measuring the value of a Moscow Apartment: a Spatial Approach to the Hedonic Pricing of Attributes

Alexandre Repkine
College of Commerce and Economics
Konkuk University
Hwayang 1 Dong
Gwangjin Gu
Seoul 143-701
Korea

Abstract

In this paper we provide a detailed exploration of the spatial effects for a large sample of apartments in the secondary market in Moscow by applying the spatial econometric techniques to the estimation of hedonic price functions with fourteen quality attributes. We find strong evidence of both spatial lag and spatial autocorrelation in our sample. The estimated values of quality attributes' 'net' contributions to the apartments' prices are robust across both the spatial model specifications and the choice of the spatial weight matrices. The fact that these values are not much different from the OLS (ML) estimates suggests that spatial effects are orthogonal to the quality characteristics making spatial estimation especially valuable for the purposes of decomposition of the apartments' price into quality *and* spatial components. One interesting finding is that an increase in the kitchen area contributes much more significantly to the apartment's price compared a marginal increase in the living area, which is reflecting the traditional role kitchen has been playing in the Russian households as a dining and communication area. House type, time needed to walk to the nearest subway station and subway time to the city center were other apartment attributes our analysis consistently identifies as important. Our study also illustrates the need for further development of the spatial econometric techniques designed for the analysis of spatial environment with both spatial lag and spatial autocorrelation effects strongly present. Finally, since this study has been actively using the open source software such as R and GeoDa, we would like to stress the importance of such packages for the efficient research process.

1. Introduction

This study is estimating hedonic price functions for the apartments market in Moscow with specific focus on the spatial heterogeneity present in the data. While there is little doubt that location is of primary importance to the housing market anywhere, in many instances, including the policy-making one, the discussion is focused on the evolution of the price of the apartments' square meter with respect to the apartments' structural and neighborhood characteristics such as the quality of high school education (Sedgley et al., 2008), air pollution (Kim et al., 2003; Habb, 2002) or the airport noise (Cohen and Coughlin, 2007). These and other studies demonstrate that these characteristics are important, yet the omission of less obvious spatial characteristics such as the average price of the neighboring apartments or the unobserved location-related characteristics may result in biased and inconsistent estimates of the individual contributions of the apartments' characteristics (Dubin, 1992).

In this study we focus on the two broadly defined spatial effects, namely, the spatial lag and spatial correlation (Anselin, 1988). Spatial lag models capture the spatial correlation structure that results from the interaction between economic agents in a specific neighborhood. In this way spatial lag models are incorporating the peer effects that otherwise cannot be observed directly. In terms of the housing market in general and in this study in particular the spatial lag model treats the *average price of the apartments* in the neighborhood as one of the determinants of the apartments' prices. Intuitively, the same apartment in terms of the observable quality characteristics would cost more in a prestigious hence more expensive neighborhood compared to a house in the backwater. Using the analogy with the time series analysis, including the spatial lag variable is similar to detrending the time series data (Anselin, 2007). Spatial correlation models, on the other hand, capture the unobserved locational effects that are shared by the neighboring units. Since these effects cannot be modeled directly as e.g. by inclusion of an additional variable as is the case with the spatial lag models, they are modeled by modifying the specification of the error term. Specifically, the error term is assumed to correlate with the error terms in the neighboring locations. In case of the housing markets such unobservable effects can be common perceptions of the future changes in the neighborhood's attractiveness in general.

In this study we are analyzing the housing market in Moscow since the latter is representing the newly born market economy enjoying impressive rates of growth in the recent years. Yet, to our knowledge there have been no studies of the hedonic housing relationships in Moscow with or without spatial effects. We demonstrate that the spatial effects are present and strong in Moscow as well even if the choice of quality characteristics in our study is in a way unlike that in the similar housing studies. Along with the obvious characteristics such as the apartments' area and the house type we include three characteristics that we believe are unique in the Russian context. One is the kitchen area that we find to be more important than the living area in terms of the effect of a marginal increase on the apartment's price.

Historically Muscovites as well as most other Russians considered the kitchen room to be not only the cooking place, but also (or even more) the place to convene with the family, discuss everyday problems and watch TV. In other words, the kitchen room in Russia plays a role similar to the one played by the dining room almost elsewhere. For example, unlike cities like Seoul where typically only the living area is reported by the numerous real estate areas the kitchen area is a must-be piece of information. In this study we (quite expectedly so) found a positive effect of a marginal increase in the kitchen area on the apartment price, but we were surprised to see how much larger it is compared to the similarly positive effect of the marginal increase in the living area.

The other two variables that are rather unique to Moscow relate to the subway. The Moscow subway, or *metro*, has been playing an outstanding role in the life of the city ever since the first subway station was built in the early 1930's. Given the stagnant economic environment during the Soviet times, individual automobiles were more of a luxury rather than a commodity of every day use[1]. Given the size of Moscow (Moscow is roughly a circle 30 kilometers in diameter) and the fact that most people were employed in the large state-owned

[1] The well-known saying during the Soviet times went: an automobile should be the means of transportation rather than a luxury. Unfortunately, obvious as it was, this never became reality until several years after the collapse of the Soviet Union in 1991.

enterprises hiring workers from all corners of the city, subway was the only rational means of commuting to and from work. Paradoxically, even the soaring number of automobiles in the city during the past ten years only exacerbated the importance of the Moscow subway since the omnipresent traffic jams and the time spent in them keep the subway the only way to get to work on time without having to get up at very early dawn. In our housing hedonic price function we include the walking time to the nearest subway station and the time needed to commute to the city center by subway from the nearest station. As expected, an additional minute in either dimension results in a decrease in the apartment's price comparable to the decrease resulting from "cutting off" one square meter of the living area.

From the methodological point of view, this paper emphasizes the importance of incorporating the spatial structure into the econometric analysis of hedonic price functions. We find that even if, in accordance with the other studies (e.g. Kim et al., 2003) the spatial lag and autocorrelation estimates are not much different from the OLS or maximum likelihood ones, some coefficients' estimates undergo significant changes compared to the OLS/ML estimates. More importantly, the intercept term magnitude is substantially revised down in case of the spatial lag model, which is evidence of the fact that a typical Moscow's apartment price has a not negligible neighborhood component.

Apart from emphasizing the importance of spatial effects in the case where geography obviously matters, like it does in the housing market, we have found out an area of possible future research in the spatial estimation techniques. Thus, even if formally the spatial lag model is preferred over the spatial autocorrelation one due to the higher value of Lagrange multiplier test statistic for the former, our statistical tests are strongly indicating the presence of *both* spatial effects in our data sample. The most we can do now given the state of art in the spatial econometrics field is to report the results for both types of models. Estimating the model that incorporates both types of spatial effects is then of obvious interest, but also a challenge since it is not clear how to separate the two effects in a single estimation procedure without introducing additional constraints on the data. In Section 3 we briefly explain why.

Finally, we gratefully acknowledge the efforts by so many people in developing

and maintaining the open source R software project (http://cran.r-project.org/) that we used together with GeoDa (another open source software) and Stata for the estimation procedures done for this paper.

2. Data

The data at our disposal comes from the Rambler Internet search engine in Russia, its real estate section (orsn.rambler.ru). Along with Yandex (www.yandex.ru) Rambler is one of the two major search engines on the Russian Internet. Its real estate section allows one to search for real estate entities according to several criteria such as the city (Moscow and St. Petersburg being the major two cities in Russia) and apartment characteristics such as the price, number of rooms, type of balcony etc.

There two major apartment types in Moscow are apartments in the newly constructed houses (*novostroyki*) and apartments in the houses that have already changed hands at least once (*vtorichnyy rynok or secondary market*). We are concentrating on the secondary market apartments since they are most representative of the Russian middle class compared e.g. to the elite apartments or stand-alone houses located in the guarded compounds. The newly constructed houses are excluded from the sample since prices for the apartments in these houses start getting quoted long before the construction of the apartment house is finished, very often getting revised either way after the house actually materializes and its apartments starts getting traded in the market.

We started with a Moscow sample of more than three thousand observations on the apartments whose size was reduced to 1125 as a result of excluding the obvious outliers and observations with missing attributes. Below is the list of attributes we have considered:

1) Living area
2) Kitchen area
3) Number of rooms in the apartment (typically one to three)
4) The floor the apartment is in (first floor is the ground floor)

5) The total number of floors in the house
6) First-last floor (a dummy)
7) Type of the apartment house (four types)
8) Distance to the nearest subway station
9) Distance to the city center
10) Bathroom type (toilet separately from the bath tub/shower)
11) Type of balcony

While the choice of the living area of the apartment as one of the attributes is rather obvious, that of the kitchen area is not. The kitchen in Russian houses is special since traditionally in the Soviet times and very often now the kitchen quite often assumes the role of a communication parlor. In that way kitchen in the Russian apartment is not just a place to cook meals, which makes it very different from kitchens in the other countries such as e.g. Korea where kitchen is exactly what it is supposed to be—a cooking place, very often not even being a separate room. This is why we believe the kitchen floor space is an important apartment attribute in Russia.

As mentioned above, the typical range of the amount of rooms in the Russian apartments is from one to three, with the four- or five-room apartments being considered as somewhat luxurious. The important characteristic of an apartment is the floor it is on (to avoid confusion, we will refer to the ground floor as the first floor). While hardly making sense in the developed part of the world, definitely back in the Soviet Union days and very often nowadays, almost nobody wanted to live on the first or the last floor, the reason being water supplies problems in these two cases. While the stories in between the first and the last floor did not experience these problems, the first and the last ones did. This is why we included a dummy that is equal to one if the apartment is located on the first or the last floor. The total amount of stories in the building is also included in the set of apartment attributes since typically buildings with more stories were also of better quality. For example, the famous Stalin high-rise buildings (also popularly known as *wedding-cake buildings* among expatriates) are of much better quality compared to the Khruschev-built five-storey houses whose main purpose around fifty years ago was to house as many families as possible.

We have five house types available in our sample: block, brick, monolith, panel and Stalin-type. Stalin-type houses were typically constructed a long time ago (first houses of this type started to appear in 1930-s) and therefore water pipes are often in a poor state or even in need of replacement. On the positive side, apartments in these houses have high ceilings, are rather spacious and are reasonably soundproof. The block- and panel-type houses are constructed according to similar technologies in the sense that they consist of the uniform elements such as blocks or panels. However, the panel-type houses do not allow for re-planning of the walls inside the apartments and have been often built during Khruschev time when quantity of housing was more of a priority compared to quality[2]. Brick houses are considered to be of better quality, but similar to the Stalin-type houses that were also constructed a long time ago. Finally, the monolith houses are now starting to dominate the housing market in Moscow and other major Russian cities. Due to the specific technology these houses allow for both almost free re-planning capabilities inside the apartment and have no seams through which sounds or even water can penetrate either from the outside or from the adjacent apartments, a problem many panel and block-type houses are infested with.

Even if Moscow has undergone visible economic transformation during the past twenty years, subway remains the major means of transportation in the city notwithstanding the rapidly increasing amount of cars in the city. For that reason, distance to the nearest subway station is an important apartment characteristic since it is directly related to the amount of time muscovites have to spend every day commuting to and from work. For the time being, there are 176 subway stations in Moscow, of which 132, or 75%, are covered by our database.

We employ two subway-related variables in our study—"distance to subway" measured as the amount of walking time in minutes to the nearest subway station and "distance to center by subway". The latter variable is measured as the smallest amount of time needed to commute from a subway station to the "Alexandrovskiy Sad" subway station, which is located literally under the Kremlin walls. We used the interactive map on http://www.metroway.ru/ in order to find that shortest time. We also believe that this way of measuring the extent

[2] The infamous *Khruschevka's* or Khruschev-time houses are notorious among Russians for low level of comfort and almost no sound isolation.

of remoteness from the city center is superior to the geographical distance measurement for the obvious reasons of traffic jams and the high degree of reliability of the subway in terms of commuting time irrespectively of the time and type of the day (e.g. working days versus holidays)[3].

Bathrooms in the Moscow apartments are either separated (the toilet bowl and the shower/bathtub are located in two separate rooms) or combined. Since traditionally apartments where the toilet seat was located in a different room than the shower/bath tub were considered to be more preferable, we expect to find a positive effect of the "separatedness" of the bathroom on the apartment price. Our balcony dummy is equal to one if the balcony is glass-insulated, also with the expected positive effect. Table 1 summarizes the data sample at our disposal.

[3] In several cases where two subway stations had exactly the same names being located on two different lines in very close proximity to each other, an average time to *Alexandrovskiy Sad* was taken (*Arbatskaya, Byelorusskaya, Kiyevskaya, Kitay-gorod, Komsomolskaya, Kuntsevskaya, Kurskaya, Novokuznetskaya, Paveletskaya, Smolenskaya, Taganskaya,*).

Table 1: Descriptive statistics of the characteristics of Moscow apartments Correlation tables

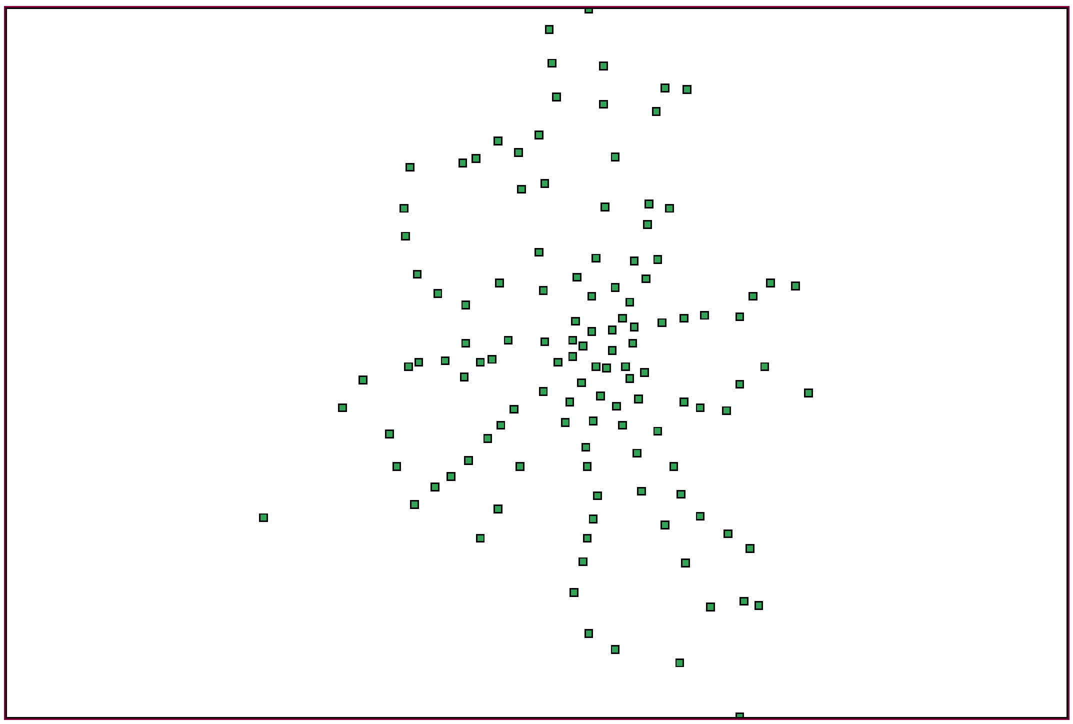
Variable Name	**Mean**	**St.Dev.**
Price (Rb, $1~25Rb)	4804880	6462290
Living area, sq.m.	46.2	27.5
Kitchen area, sq.m.	10.23	4.92
Floor	7	4.74
Total number of floors	12	5.9
Number of rooms	2.6	1.6
Time to subway, min walk	15	12.3
Time to center, min subway ride	25	9.5
First/last floor apartment, number of observations	206	
Separated toilet / bath tub, number of observations	889	
Insulated veranda, number of observations	514	
Block, number of observations	46	
Brick, number of observations	341	
Monolit, number of observations	166	
Stalin, number of observations	87	

Total number of observations: 1125

In our sample, the average price of a Moscow apartment is in the area of $200,000, mostly represented by the mostly widespread panel houses. In this way in our study we are concentrating on the middle class segment of the housing market excluding luxury apartments and the newly built apartment houses. Kitchen on average constitutes around a quarter of the total living area with the latter averaging 46 square meters. An average apartment in our sample is located in the middle store of the typical multi-storey building that on average has twelve floors. Two to three rooms would represent a typical Moscow apartment according to our sample. A twenty minutes subway ride would be enough for a muscovite to get to the city center, but he or she would need an additional fifteen minutes to walk to that subway station. Around one-fifth of our apartments are unfortunate to be located on the first or the last floor, which we expect to produce a negative influence on the apartment price. Around 80% of the apartments are boasting separated toilet seat and the bath tub, representing the traditional preference for such separation. Finally, around one-half of our apartments are enjoying an insulated veranda (winters in Moscow can be very cold!)

In general, we believe our sample is representative both in terms of the quality characteristics of a typical Moscow middle class apartment as well as geographically. The figure below demonstrates the geographical distribution of the apartments in our sample (the map is representing apartment centroids taken to be the nearest subway stations).

Figure 1: Geographical Location of the Sample Apartments



3. Theoretical Framework

The starting point of our analysis is the hedonic price function of a multi-attribute apartment formulated in a seminal paper by Rosen (1974). The price of a house in the hedonic price function framework is a function of its quality characteristics as well as of those of the neighborhood: $P = f(X_s, X_n) + \varepsilon$, where X_s is the vector of the apartment's characteristics such as the living area or the number of rooms and X_n is the vector of characteristics of the neighborhood (in our study distance to the nearest subway station and subway commuting time to the city center). All in all, we employ fourteen characteristics in this study.

Consumers maximize their utility that is a function of the apartment composite good $X = (X_s, X_n)$ subject to the constraint $I = C + P(X)$ where I is income,

C is a numeraire commodity and $P(X)$ is the price of apartment $X=(X_s, X_n)$. The shadow price of the apartment's attributes can be then shown to be equal to the ratio of the marginal utility of the apartment attributes to that of the numeraire good: $\frac{\partial p}{\partial z}=\frac{\partial U/\partial z}{\partial U/\partial p}$. We can then estimate those shadow prices by specifying an empirical form of the hedonic price function.

As we mentioned in the introduction section, there are two ways in which spatial effects can be added to the standard hedonic price function specification. The spatial lag model accounts for the effects of the prices of neighboring apartments by specifically adding the spatially lagged variable to the specification equation: $P=f(X_s, X_n, WP)$, where W is the spatial weights matrix.

The spatial weights matrix W is in general a sparse matrix with non-zero elements accounting for the neighboring units. For example, $w_{2,5}\neq 0$ means that the second apartment in the sample has the fifth apartment as one of its members. Theoretically there are not many indications as to which form of matrix W is most appropriate. The simplest form of W is the one whose elements are equal to unity if the corresponding apartment is neighboring the one representing a specific row, and zero otherwise. In that case the rows of W are often standardized so that the elements' sum in each row may be equal to one. We can then interpret the spatially lagged variable WP as the average price of the neighboring apartments. This is one of the approaches we are undertaking in this paper. This is also the approach that appears to be used most frequently (see e.g. Cressie, 1993). Another approach is to postulate that each apartment in the sample has a specific number of its closest neighbors, which will result in a different W. The problem is the choice of the appropriate number of closest neighbors. We deal with this problem in the next section. Our results are robust across the choice of the spatial weights matrices qualitatively, but not quantitatively.

The second way in which spatial effects can be taken into account is by means of the spatially correlated errors. In that case the set of independent variables remains the same, but the error process is described in a different way as follows:

$$P = f\left(X_s, X_n\right) + \varepsilon$$

$$\varepsilon = \lambda W \varepsilon + u ,$$

where u is the conventional error term. Typically such a representation is representing the environment in which neighboring apartments share a particular characteristic, but the latter is not directly observable. It can be, for example, the criminal situation or the air quality (neither one is available in our sample).

The estimation for both types of the spatial dependence described above is typically done by the maximum likelihood. The problem is that if the residuals' analysis indicates that the latter are not distributed normally, maximum likelihood estimates are not appropriate. Applying the two stage least squares estimation with lagged independent variables as instruments is robust to non-normality of errors and results in unbiased and consistent estimates (Kim et al., 2003). We do discover non-normality in the residuals, so we report both maximum likelihood and two stage least squares estimates. If the heteroskedasticity in the residuals caused by either remaining spatial dependence or any other factors is not removed after the two stage estimation, the heteroskedasticity-robust version of the two stage estimation procedure should be applied. Since in our case such remaining heteroskedasticity is indeed the issue, we apply precisely this procedure in order to obtain more efficient estimates.

As we mentioned already, the choice between spatial lag and spatial correlation model is made on the basis of the Lagrange multiplier test statistics and the "combination" tests that test for either form of spatial dependence *in the presence* of the other type of the one. One problem that could not be resolved in the framework of this study is that both forms of spatial dependence were strongly suggested by the statistical testing.

The not so minor issue here is that it is not clear how to proceed with the estimation of the two effects simultaneously. Indeed, the model in question (we assume the most general form in which the spatial lag structure is described by a spatial weights matrix that is different from the one characterizing spatial correlation)

$$\begin{cases} y = \rho W_1 y + X\beta + \varepsilon \\ \varepsilon = \lambda W_2 \varepsilon + u \end{cases}$$, where u is normal i.i.d. (1)

Clearly, (1) can be re-written as:

$$\left(I - \lambda W_2\right)\left(I - \rho W_1\right) y = \left(I - \lambda W_2\right) X\beta + u \qquad (2)$$

Estimating (2) is rather problematic since it is not immediately clear how to separately identify the effects of λ and ρ.

Resolving the issue of simultaneous estimation of two types of spatial dependence and interpretation of the results is beyond the scope of this study, but we believe it is an interesting area of future theoretical research in the area of spatial econometrics. In this study we report the results of the estimation of both types of spatial dependence.

4. Empirical Results

Choosing hedonic price function specification

Our first step will be to determine whether spatial effects are indeed present in our sample so that we know the OLS estimates of the hedonic price equation are likely to produce biased and inconsistent results. Before doing spatial diagnostics on the OLS regression, though, we first need to identify the appropriate functional form for the hedonic equation. Unfortunately, economic theory provides us with little guidance as to the functional form of hedonic price equations (Halvorsen, 1981). Since to our knowledge the spatial effects have not yet been properly incorporated into the flexible functional forms (e.g. Box-Cox transformation) our choice is between double-log, semi-log, linear-log and linear-linear specifications with respect to the dependent and independent variables, respectively.

We choose between the set of logged and not logged independent variables by running the regression collinearity diagnostic procedure developed in Belsley et al. (1980) that are based on computing the conditioning number of the matrix of the independent variables. The conditioning number is computed as a ratio of the largest eigenvalue of the matrix of independent variables to the lowest one. Values of this number above 30 indicate an unacceptable degree of multicollinearity in the independent variables. We estimate the conditioning number for our sample to be 45.76 in case of the logged values and 14.41 in case of the original values. We are therefore left with the semi-log and linear-linear specifications. Since the OLS estimates of the former produce a much higher value of R-squared (76%) compared to the latter (46%), we choose the semi-log specification of the hedonic price function as a basis for our further analysis. Table 2 below presents OLS estimates of the semi-log specification. The rightmost column of the table translates OLS estimates into the percentage increase in the apartments' price due to a unit increase in the value of one of its attributes.

Table 2: OLS estimates of the hedonic price equation

Dependent variable: (log) apartment price, Rb

	Semi-Log Specification	**Percentage Increase in Price to a Unit Change**
Constant	12.07 (0.000)	
Apartment Characteristics		
Living area, sq.m.	0.007 (0.000)	0.7%
Kitchen area, sq.m.	0.03 (0.000)	3.05%
Floor	0.002 (0.4)	0.2%
Total Floors	0.004 (0.14)	0.4%
Number of Rooms	0.17 (0.000)	18.53%
Neighborhood Characteristics		
Time to Subway (minutes)	-0.01 (0.000)	-1%
Time to Center (by subway, minutes)	-0.02 (0.000)	-1.98%
Apartment Structural Dummies		
First/Last Floor Dummy	-0.11 (0.000)	-10.42%
Toilet and Bath Separated Dummy	0.008 (0.76)	0.8%
Veranda Dummy	0.05 (0.03)	5.13%
House Type Dummies		
Block	-0.07 (0.2)	-6.76%
Brick	0.17 (0.000)	18.53%
Monolith	0.18 (0.000)	19.72%
Stalin	0.18 (0.000)	19.72%
Regression Diagnostics		
R-squared	0.756	

Adjusted R-squared	0.752	
No. Observations	1125	
F-statistic	245.22	

Discussion of the OLS estimation results

As mentioned already, the semi-log specification provides a more reasonable rate of fit (75%) compared to the linear-linear specification, so our discussion in this section will be based on the semi-log OLS results.

Both living and kitchen area command predictable positive and statistically significant coefficients. However, the effect of a marginal increase in the kitchen area far outweighs that of an increase in the living area, reflecting the important role kitchen has been traditionally playing in the Russian homes: an additional square meter in the kitchen raises the apartment price by 3%, while that in the living area only does that by 0.7%. The number of rooms is also producing a positive and significant effect on the apartment's price with one additional room adding 20% to the apartment's price.

With respect to the floor characteristics, only the first-last floor dummy comes out statistically significant with these apartments losing 10% in their value, reflecting the specific water supply system characteristics of the Russian construction. As long as the apartment in not on the first or last floor, the households appear not to care about the exact number of their floor or the total number of floors in the house. Neither do Moscow households appear to care whether the toilet is separated from the bathroom. However, the absence of a veranda turns out to be a significant apartment characteristic for them at 5% of the apartment's price.

With respect to the house type, brick, monolith and the Stalin-type buildings command a premium over the *Khruschevka* and block-type buildings at nearly 20% of the price reflecting the low-quality of sound isolation and general lack of prestige for the latter two types of housing. Finally, the amount of time Muscovites spend getting to the nearest subway station and the time they need to go to the city center predictably come out important with one more minute diminishing the apartment's price by one and two percent, respectively.

So far the OLS results have not been counterintuitive except maybe for the fact that the separate toilet and bath do not seem to produce any significant effect on the apartments' price. However, as is well known, the OLS estimates are inconsistent in the presence of spatial errors while in case spatial lags are strongly present, the OLS estimates also become biased. For that reason we take these estimates only as preliminary evidence and proceed with the analysis of possible spatial effects in our sample.

Testing for heteroskedasticity of the residuals

Both in case of the spatial lag and spatial error models the Gauss-Markov assumptions on the error terms are going to be violated in the sense that the latter are no longer homoskedastic. Indeed, in case of the spatial lag model heteroskedasticity in the residuals will be caused by the omitted variable (W*LNP), while in case of the spatial error model the error variance varies according to location by definition. For that reason we start by finding evidence of heteroskedasticity in the residuals and then proceed with the more specific tests for spatial effects.

Both the Breusch-Pagan (Cook-Weisberg) test for multiplicative heteroskedasticity and Szroeter test for homoskedasticity against the alternative that the residual variances are monotonically increasing in the independent variables strongly suggest the presence of serious misspecifiation problems. Table 3 below presents estimation diagnostics for these tests.

Table 3: Breusch-Pagan and Szroeter tests for heteroskedasticity in OLS residuals

Breusch-Pagan test	**Chi2 (14)**	**p-value**
All 14 right-hand side variables	787.47	0.0000
Szroeter test		
Living	159	0.0000
Kitchen	218	0.0000
Floor	47	0.0000
Total floors	8.6	0.0033
Rooms	62.9	0.0000
Time to subway	6.4	0.0114
Time to center	33.5	0.0000
First/Last Floor Dummy	9.83	0.0017
Toilet and Bath Separated Dummy	4.7	0.0302
Veranda Dummy	26.61	0.0000
Block	12.8	0.0003
Brick	20.5	0.0000
Monolith	4.24	0.0394
Stalin	4.6	0.0322

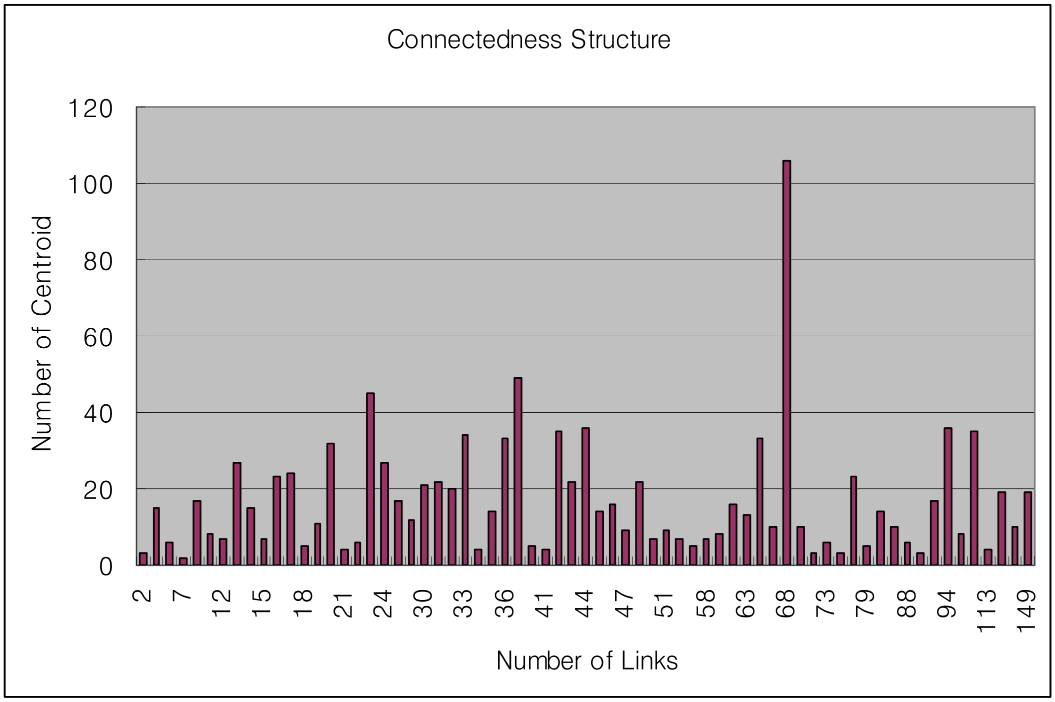
Now that heteroskedasticity in the residuals indicates the possible presence of spatial effects, we proceed with the more specific tests. In particular, we perform a series of tests in order to see whether the spatial effects are indeed present and what nature these effects are of. Since all of these tests are based on a specific distance weight matrix, we first discuss the weight matrix type most appropriate for our purposes.

Choosing the weights

Most generally, there are two broad approaches to defining the weights in the spatial weight matrix *W* (LeSage and Hill, 2004). One is based on the geographical (or any other conceptual) distance between the objects (apartments in our case), while the other is more suitable for the samples where any reasonable minimum geographical distance results in too many “islands”—objects with no neighbors that must be therefore dropped out of the analysis. Since visual inspection of our sample (Figure 1) does not readily provide us with an insight as to which one of the two weight schemes is the most appropriate, we proceed as follows.

We begin by computing the *W* matrix based on the geographical distance with the minimum distance within which two apartments are considered to be neighbors to be the one for which there are no “islands” in the sample. In this way we hope to avoid eliminating observations out of our analysis as well as driving the number of neighbors unrealistically high by choosing the distance to be such that everyone is everyone’s neighbor. Our distance-based matrix constructed in this way contains three least connected regions with only two links, while the 19 most connected regions are connected to 149 neighbors. The average number of links is 50 with the share of non-zero links equal to 4.5%. Figure 2 below presents a distribution of apartment centroids by the number of links.

Figure 2: The Distribution of Apartment Centroids by the Number of Neighbors



We do not observe any well-defined mode in the distribution above, however, we can hardly characterize it as uniform either. The advantage of matrix *W* represented by this distribution is that it has no zero-only rows, or equivalently, every apartment centroid has at least one neighbor. This is a desirable property since we do not have to keep track of the missing observations (the "island" ones) when performing our spatial regression analysis. Finally, since our distance-based matrix is a contiguity one, it is also symmetric, which we checked for.

As we mentioned, another approach to constructing the spatial weights matrix is to compute the weights on the basis of *k-nearest neighbors.* One of the advantages of this approach is that the distance-based approach applied above normally results in a rather unbalanced connectedness structure due to e.g. clustering. In order to carry out our analysis on a more balanced connectedness structure, we now postulate that each apartment in our sample has at least *k* neighbors (which it does by definition since the number of units in our sample is far larger than five). The clear-cut advantage of this approach is that the "islands" problem is automatically removed, while the overall connectivity structure becomes balanced, although in most cases not symmetric. The biggest problem is the "right" number of the nearest neighbors. Since there is no immediately available theoretical guidance on that number, we choose it to be the minimum number of links calculated for the distance-based spatial weights matrix, namely, two. This number is one notch away from the default number of neighbors in e.g. a popular program GeoDa that we used to calculate these weights.

The percentage of non-zero weights in the k-neighbors based matrix is 0.17, which is much lower compared to the distance-based one. By construction, every apartment centroid has exactly two neighbors, but as neighbors are not defined in terms of contiguity, the spatial weights matrix is not symmetric. To reiterate, using the k-nearest neighbors based matrix allows us to avoid the problem of unbalanced connectivity structure at the expense of diminished importance of actual geographical distance.

In what follows we are going to use these two matrices in order to specifically test for the two types of spatial dependence in our data and then use the results of these tests in choosing the appropriate type(s) of the spatial regression model.

Testing for spatial dependence

The following table summarizes the results of five Largrange multiplier tests each one regarding a specific aspect of spatial dependence.

Table 4: Lagrange multiplier tests for spatial dependence

	LM test statistic	**p-value**
Simple error dependence		
Distance-based W	272	0.0000
Nearest neighbors W	161	0.0000
Simple spatially lagged variable		
Distance-based W	228	0.0000
Nearest neighbors W	143	0.0000
Error dependence in the presence of spatial lag		
Distance-based W	115	0.0000
Nearest neighbors W	54	0.0000
Spatial lag in the presence of error dependence		
Distance-based W	71	0.0000
Nearest neighbors W	35	0.0000
Joint test on error dependence *and* spatial lag		
Distance-based W	343	0.0000
Nearest neighbors W	197	0.0000

The first two tests in the table (simple error and simple spatial lag tests) are what their names suggest: the assumption is the data are either characterized by spatial autocorrelation or a spatial lag. The next pair of the tests are essentially elucidating the same issue except this time the possible presence of the other type spatial dependence is taken into account so that these tests are robust to the presence of other type spatial effects. The final test in the table seeks to answer the question whether the most general spatial model incorporating both spatial autocorrelation and spatial lag is in fact governing the data. Finally, we have performed each one of the five tests on two spatial weight matrices whose construction was described above: the one based on geographical distance and the one based on *k* nearest neighbors (*k* was chosen to be equal to two for consistency reasons since this is the minimal number of neighbors according to the weights matrix constructed on the basis of the distance threshold).

First off, the p-value for each one of the ten tests performed was found to be equal to the machine zero (1e-16), which leaves little doubt in the fact that the spatial effects are indeed strongly present. At the same time, it is not immediately clear whether the spatial lag specification should be preferred to the spatial autocorrelation one. On the one hand, irrespectively of the spatial weights matrix *W* or the extent of the robustness of the tests the Lagrange multiplier test statistics are greater for the spatial lag specification compared to the case of the spatial error one, so that the spatial lag model appears to be more appropriate. On the other hand, the all-comprehensive test for the simultaneous presence of the two effects results in the highest value of the test statistic suggesting both spatial effects are strongly present. Besides, given the extremely low p-values for these tests one can hardly argue that the spatial lag specification indeed represents the data more adequately relative to the spatial error one. For that reason we are presenting the results of empirical estimations for the spatial error and the spatial lag model separately keeping in mind that the LM test statistics came out larger in case of the spatial lag model.

Spatial error/spatial lag estimation results

Table 5 below presents the results of our estimates for the spatial lag (first three columns) and the spatial error (the last column) specifications. The two columns in the middle represent the results of two-stage estimation of the spatial lag model that deals with potential endogeneity of the spatially lagged variable *Wy*. Two-stage estimation is also robust to the non-normality of errors, which in our case is confirmed by the large values of the Jarque-Bera statistic for their normality test. The spatially lagged housing and neighborhood characteristics (e.g. the independent variables in the hedonic model without spatial effects) are used as instruments for the spatially lagged dependent variable. The robust version of this two-stage estimation (the third column) applies heteroskedasticity correction to the coefficient estimates and their covariances in order to account for the effects of possibly remaining heteroskedasticity. Table 6 presents the same set of estimates for the case of the spatial weight matrix constructed on the basis of *k-nearest neighbors* (two neighbors in our case).

Table 5: Spatial Models Estimation in Case of the Distance-Based Spatial Weights Matrix

	Spatial Lag Model Maximum Likelihood Estimation	Two Stage Spatial Lag Estimation	Two Stage Spatial Lag Estimation Robust to Heteroskedasticity in Residuals	Spatial Error Model
Constant	6.53 (0.0000)	6.25 (0.0000)	6.53 (0.0000)	12.07 (0.0000)
Apartment Characteristics				
Living area, sq.m.	0.007 (0.0000)	0.007 (0.0000)	0.008 (0.0000)	0.006 (0.0000)
Kitchen area, sq.m.	0.03 (0.0000)	0.03 (0.0000)	0.03 (0.0000)	0.03 (0.0000)
Floor	0.003 (0.22)	0.003 (0.22)	0.002 (0.4)	0.002 (0.47)
Total Floors	0.004 (0.08)	0.004 (0.08)	0.005 (0.04)	0.007 (0.006)
Number of Rooms	0.16 (0.0000)	0.16 (0.0000)	0.14 (0.0002)	0.17 (0.0000)
Neighborhood Characteristics				
Time to Subway (minutes)	-0.008 (0.0000)	-0.008 (0.0000)	-0.008 (0.0000)	-0.007 (0.0000)
Time to Center (by subway, minutes)	-0.006 (0.0000)	-0.006 (0.0002)	-0.004 (0.0002)	-0.02 (0.0000)
Apartment Structural Dummies				
First/Last Floor Dummy	-0.1 (0.0000)	-0.1 (0.0000)	-0.09 (0.0000)	-0.1 (0.0000)
Toilet and Bath Separated Dummy	0.01 (0.61)	0.01 (0.60)	0.02 (0.34)	0.009 (0.72)
Veranda Dummy	0.06 (0.004)	0.06 (0.004)	0.05 (0.01)	0.05 (0.01)
House Type Dummies				

Block	-0.06 (0.25)	-0.06 (0.26)	-0.05 (0.06)	-0.07 (0.16)
Brick	0.15 (0.0000)	0.15 (0.0000)	0.16 (0.0000)	0.15 (0.0000)
Monolith	0.18 (0.0000)	0.18 (0.0000)	0.16 (0.001)	0.20 (0.0000)
Stalin	0.17 (0.0000)	0.17 (0.0000)	0.17 (0.0000)	0.19 (0.0000)
Estimation Diagnostics				
Pseudo R-square	77.94%	78.3%	78.1%	78.44%
No. Observations	1125	1125	1125	1125
Heteroskedasticity in residuals (Breusch-Pagan test)	64 (0.0000)			56 (0.0000)
Rho (Lambda)	0.41 (0.0000)	0.43 (0.0000)	0.41 (0.0000)	0.64 (0.0000)
LM test for residual autocorrelation	6 (0.01)			

Table 6: Spatial Models Estimation in Case of the Spatial Weights Matrix Based on the Two Nearest Neighbors

	Spatial Lag Model Maximum Likelihood Estimation	Two Stage Spatial Lag Estimation	Two Stage Spatial Lag Estimation Robust to Heteroskedasticity in Residuals	Spatial Error Model
Constant	9.99 (0.0000)	10.03 (0.0000)	10.38 (0.0000)	12.15 (0.0000)
Apartment Characteristics				
Living area, sq.m.	0.007 (0.0000)	0.007 (0.0000)	0.01 (0.0000)	0.006 (0.0000)
Kitchen area, sq.m.	0.03 (0.0000)	0.03 (0.0000)	0.03 (0.0000)	0.03 (0.0000)
Floor	0.003 (0.32)	0.003 (0.32)	0.002 (0.43)	0.002 (0.37)
Total Floors	0.003 (0.17)	0.003 (0.18)	0.003 (0.25)	0.004 (0.11)
Number of Rooms	0.16 (0.0000)	0.16 (0.0000)	0.09 (0.005)	0.17 (0.0000)
Neighborhood Characteristics				
Time to Subway (minutes)	-0.009 (0.0000)	-0.009 (0.0000)	-0.009 (0.0000)	-0.009 (0.0000)
Time to Center (by subway, minutes)	-0.013 (0.0000)	-0.013 (0.0000)	-0.01 (0.0000)	-0.019 (0.0000)
Apartment Structural Dummies				
First/Last Floor Dummy	-0.11 (0.0000)	-0.11 (0.0000)	-0.09 (0.0000)	-0.10 (0.0000)
Toilet and Bath Separated Dummy	0.008 (0.74)	0.008 (0.74)	0.03 (0.29)	0.002 (0.93)
Veranda Dummy	0.05 (0.01)	0.05 (0.01)	0.04 (0.05)	0.05 (0.02)
House Type Dummies				

Block	-0.08 (0.12)	-0.08 (0.12)	-0.07 (0.007)	-0.08 (0.13)
Brick	0.15 (0.0000)	0.15 (0.0000)	0.13 (0.0000)	0.16 (0.0000)
Monolith	0.19 (0.0000)	0.19 (0.0000)	0.16 (0.0010)	0.20 (0.0000)
Stalin	0.19 (0.0000)	0.20 (0.0000)	0.17 (0.0000)	0.19 (0.0000)
Estimation Diagnostics				
Pseudo R-square	77.93%	77.92%	77.1%	77.55%
No. Observations	1125	1125	1125	1125
Heteroskedasticity in residuals (Breusch-Pagan test)	70.23 (0.0000)			78.51 (0.0000)
Rho (Lambda)	0.16 (0.0000)	0.16 (0.0000)	0.13 (0.0000)	0.20 (0.0000)
LM test for residual autocorrelation	41.39 (0.0000)			

As Tables 5 and 6 demonstrate, in case the spatial weights matrix is defined in terms of the *k*-nearest neighbors, the spatial effects both in spatial lag and spatial error model are estimated to be lower. For example, parameter *rho* and *lambda* would be 0.41 and 0.64 in case of the geographical distance-based weights, while in case of the *k*-nearest neighbors specification these coefficients would be equal to 0.16 and 0.20, respectively.

One conclusion that remains valid irrespectively of the spatial model specification or the estimation method is that spatial effects are strongly present in our sample. Since the presence of both spatial lag and error patterns is strongly suggested by the statistical tests with no immediately known way of giving preference to one or the other, the following discussion will be based on the estimates for both kinds of models[4].

[4] Formally, the spatial lag model should be preferred over the spatial autocorrelation because the Lagrange multiplier test statistic is estimated to be larger for the former versus the latter for both types of the spatial weights matrices. However, the Lagrange

The pseudo R-squared are reasonable at little less than 80% in every specification. However, the high values of Breusch-Pagan statistic for the heteroskedasticity tests as well as the high values of spatial error statistic performed on the spatial lag model residuals suggest that the spatial dependency in our data has not been fully captured by either one of the specifications that we tried to estimate. We take this result as evidence of lack of the theoretical basis for estimating the models in which both spatial lag and spatial error effects are present. Since to our knowledge there is no such theoretical framework currently, we found it best to estimate the two types of spatial effects separately.

Compared to the OLS applying the "space-conscious" estimation procedures resulted in a slight improvement in the R-squared and sometimes a revision in the individual effects of the apartments' characteristics. Thus, according to our OLS estimates, one additional minute of walking time to the subway would decrease the apartment's price by 1%, while the same increase in time to the Moscow center would decrease this price by twice as much. In our spatial estimations time retains its statistically significant negative influence on the price, however, the size of the effect decreases by one-third in case of the time to subway station and by more than three times in case of the time to center. In particular, a one minute increase in the time needed to go to the city center, the apartment's price is estimated to decrease by 0.6% in case of the spatial lag model with the distance-based spatial weights matrix. The fact that most estimated coefficients are not much different from their OLS counterparts suggests that the spatial characteristics are orthogonal to the quality characteristics of the apartments (so that the omitted variable bias does not show up in the coefficients' estimates). However, the much lower intercept value in case of the spatial models estimation is demonstrating the value of spatial econometric analysis for decomposition of the apartment price into quality and spatial components.

The signs, statistical significance and the magnitude of the individual coefficients do not appear to differ much across the choice of the spatial dependence model, the choice of estimation technique (e.g. maximum

multiplier tests are uniformly suggesting the presence of both type of spatial effects.

likelihood estimation or the two-stage maximization) or the kind of the spatial weight matrix. The only exception seems to be the effect of the time to center variable that comes out larger in case of the spatial error model for both kinds of spatial weights matrices. It is also this characteristic for which the difference with OLS estimates is the largest compared to other coefficients. Similar, although less pronounced, reduction in the magnitude of the coefficient in case of the spatial models estimation is produced in case of the time to the subway station. It is worthwhile noting that these are the only two independent variables that have explicitly to do with the geographical location of the apartments. In that case some reduction in the value of these coefficients would be expected since part of the spatial structure in the data has been already incorporated econometrically.

5. Conclusion

In this paper we estimated hedonic price functions that take account of spatial dependence for a large sample of Moscow apartments. Having carefully tested for the presence of spatial dependence effects, we found both spatial lag and spatial correlation effects present in our data. Applying the log-linear specification of the hedonic price function that we chose according to the correlation conditioning number tests as well as the goodness of fit, we estimated hedonic relationship according to the two alternative types of spatial weights matrices (distance-based and k-nearest neighbor). For neither type of the matrices estimating the spatial lag or spatial autocorrelation effect failed to remove spatial dependence in the residuals, which was demonstrated by the robust spatial lag / spatial correlation tests. Even if formally according to the Lagrange multiplier test the spatial lag model is preferred to the spatial correlation one, strong evidence of the simultaneous presence of both spatial effects made us decide to present both sets of estimates. Since at the present stage and to our knowledge, no econometric tools make it possible to estimate and adequately interpret the model that includes both types of spatial dependence effects, we leave this exercise for our further research.

For each type of the spatial effect we estimate hedonic price functions with fourteen attributes gradually relaxing the assumptions of the normality of errors and taking account of the remaining heteroskedasticity in the errors even after the spatial dependency effects have been incorporated into our estimation procedure. Surprisingly, the estimated coefficients are much similar to the OLS/ML estimates of the “space-unaware” specification except for one attribute (the time to center one in case of the spatial autocorrelation) leading us to conclude that in our sample the spatial component of the market price of an apartment is orthogonal to the quality attributes’ one. In particular, including the spatially lagged variable into the hedonic price function significantly reduces the value of the OLS/ML intercept even if it fails to drastically improve the value of the (pseudo) R-square. In this way, the spatial lag model not only takes care of the possible bias and inefficiency in the OLS estimates (not a problem in our case) but it creates grounds for explicitly estimating part of the apartments’ price accounted for by nothing else but the geographical location.

In choosing our hedonic price function specification, we took account of the cultural idiosyncrasies pertinent to Moscow and Russia in general. Thus, we included the kitchen room area and two subway-related variables into the set of otherwise obvious quality attributes into our specification. The kitchen room has been traditionally viewed by most Russians as a convention place for the family, making it an important attribute in the case of Moscow apartments. The exceptional role of the subway in Muscovites' everyday commuting to and from work necessitated the choice of walking time to the nearest subway station and the subway time to Moscow center as two other quality attributes. All three attributes proved to be highly significant and had expected signs. One interesting finding was that a marginal increase in the kitchen room area affected the price much more strongly compared to a comparable increase in the living area.

In performing this study, we have largely relied on the free open-source R software for spatial regression estimation and GeoDa for construction of spatial weights. We hope that this paper would serve as one other reason why open-source software is so important for the academic community and research.

References

Anselin, L., 2007, Spatial Regression, mimeo, Department of Geography, University of Illinois.

Anselin, L., 1998, *Spatial Econometrics: Methods and Models,* Kluwer Academic Publishers, Dordrecht, Netherlands.

Cohen, J.P., and Coughlin, C.C., 2007, Spatial hedonic models of airport noise, proximity and housing prices, *Federal Reserve Bank of St. Louis working paper no. 2006-026C*.

Cressie, N.A.C., 1993, Statistics of Spatial Data, Wiley, New York.

Dubin, R.A., 1992, Spatial autocorrelation and neighborhood quality, *Regional Science and Urban Economics*, **22**, 432-452

Habb, T.C., McConnell, K.E., 2002, Valuing environmental and natural resources: the econometrics of non-market valuation. Edward Elgar, Cheltenham, UK.

Halvorsen, R., and Pollakowski, H.O., 1981, Choice of functional form for hedonic price equations, *Journal of Urban Economics,* **10**, 37-49

Kim, Ch.W., et al., 2003, Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach, *Journal of Environmental Economics and Management*, **45**, 24-39

LeSage, J.P., and Hill, R.C. (Eds.), 2004, *Spatial and Spatiotemporal Econometrics*, Elsevier Science.

Rosen, S., "Hedonic Prices and Implicit Markets : Product Differentiation in Pure Competition", *Journal of Political Economy,* **Vol. 92**, 1974

Sedgley, N.H., Williams, N.A., and Derrick, F.W., 2008, The effect of educational test scores on house prices in a model with spatial dependence, *Journal of Housing Economics,* **17**, 2, 191-200.