Spatial Econometric Approaches to Estimating Hedonic Property Value Models

Li Wang and Richard Ready

Abstract: The inclusion of spatial correlation of house price in hedonic pricing model may produce better marginal implicit price estimate(s) of the environmental variable(s) of interest. Most applications where spatial econometric models are applied to the estimation of a hedonic property value model have used either a spatial lag model or a spatial autoregressive (SAR) error model. Three issues regarding the specification of a spatial hedonic pricing model are considered. First, we argue that the spatial error components (SEC) model is more theoretically intuitive and appealing for modeling house prices. Second, we question the “convention” of row-standardizing the spatial weights matrix. Third, we explore whether the choice of spatial model is important, empirically, using a large house sale dataset that includes measures of proximity to landfills. With one exception, estimated marginal implicit prices are fairly robust across all models. However, the results show that if house prices are spatially correlated, use of an inappropriate spatial model can lead to marginal implicit price estimates that are less reliable than those from OLS regression.

Key Words: Hedonic house prices, spatial correlation, spatial lag model, SAR error model, SEC model
1. Introduction

It is widely observed in real estate data that the sale price of a house at one location is similar to that of a house located nearby for reasons other than those explicitly incorporated in the hedonic model. This is an important observation because the marginal implicit price (MIP) estimates of some environmental variables of interest, such as proximity to undesirable land use, from the traditional hedonic analysis may be biased and/or inefficient, if the spatial dimension of the data is ignored. Spatial econometrics is a tool that would remedy this problem. However, different spatial specifications exist. This paper explores which spatial specification may be theoretically more appropriate in the context of house price and how sensitive the estimation result is to the choice of spatial model.

The observed spatial correlation of house price occurs for several reasons. First, a seller of one house located very close to another similar house that was recently sold may set the selling price according to that of the neighboring house. This would result in a “herd effect”, where price expectations are formed based on neighboring values. Similarly, in cases where the selling prices are set by real estate professionals, recent local house sales, or “comparables”, are given strong weight when the asking price is set. Second, the conventional “wisdom” among real estate professionals is that the cheapest house on the block appreciates faster than more expensive neighbors so that nearby prices will tend to compress over time. Third, homeowners may receive positive utility from living close to nicer (high-valued) houses, with the result that a house’s price would be correlated with surrounding house prices. A common spatial econometric model that is motivated by any of these three processes is the spatial lag model. In the spatial lag model, the price of a given house depends on its attributes as well as on the prices of neighboring properties. If house prices are generated by a spatial lag structure, OLS
regression will give biased estimates of the parameters of the hedonic price regression (Anselin, 1988).

However, there are other processes that can result in spatially correlated house prices. For example, if there exists an omitted (unmeasured) variable that influences house prices and that varies spatially, houses located near each other will tend to have similar values for that unmeasured variable, and thus tend to have correlated prices. This process motivates a different class of spatial econometric models that allows for spatial correlation in the regression errors. A commonly used example is the spatial autoregressive (SAR) error model. If house prices are generated by a process with spatially correlated errors, OLS regression will tend to give unbiased but inefficient parameter estimates, and OLS t-statistics will be biased (Anselin, 1988).

There are many recent applications of spatial econometric approaches in hedonic pricing studies in the context of housing markets. Their findings demonstrate the consequences of aspatial specification and suggest that an explicit spatial hedonic specification is beneficial (although not always required for every hedonic analysis). Pace and Gilley (1997) found that the SAR error model improved the overall prediction of house price (a 44% reduction of the errors relative to the OLS). Can and Megbolugbe (1997) found that a model with a spatially dependent variable not only increased the explanatory power of their model, reflected in a higher R², but also addressed to some extent the problem of omitted house structure variables. Wilhelmsson (2002) found that there exists some spatial effects in house price data and that both the spatial lag model and the SAR error model explain more of the variation in price than does OLS. In a study valuing air quality, Kim et al (2003) report that the spatial lag model is favored over the SAR model, and that OLS overestimates the effect of air quality on house price in Seoul, Korea in the presence of spatial lag dependence.
The spatial lag and SAR error models are the two most commonly used models in hedonic pricing studies, in part because routines are available for their estimation, e.g., the Spatial Econometrics Toolbox for Matlab by LeSage and GeoDa by Anselin. While some authors have explored which of these two models might provide a better fit to a particular dataset, there has been less exploration of the sensitivity of externality estimates to the choice of spatial model, and even less reflection on which models might be theoretically more appropriate for hedonic analysis of house price. Another often ignored issue is the “conventional” use of a row-standardized spatial weights matrix. Row-standardization is appealing in part because the associated spatial parameter has a clear interpretation as a measure of spatial dependence (autocorrelation). This makes the spatial parameter comparable between models. However, as Bell and Bockstael (2000) argued, row-standardization changes the assumed spatial structure of the sample data and so the intended “economic” relationship among observations.

The purposes of this paper are to 1) explore the theoretical implications of alternative spatial econometric models, and assess the degree to which they are consistent with our understanding of house price formation; 2) explore the theoretical implications of using a row-standardized spatial weights matrix in the context of hedonic price models; and 3) investigate the sensitivity of marginal implicit price estimates for an environmental disamenity to alternative model specifications. In addition to two most-commonly-used spatial econometric models, we consider a third model, the spatial error components (SEC) model described by Kelejian and Robinson (1993, 1995).

The remainder of the paper is organized as follows. In section 2, the three spatial econometric models are presented, and the theoretical implications of each model are explored, and their suitability for house price data is discussed. In section 3, the consequences of row
standardization of the spatial weights matrix are considered for each model. In section 4, the
three models are estimated using a dataset on house sales in Berks County, PA, with particular
attention to estimation of the marginal implicit price of proximity to a local disamenity, landfills.
Section 5 concludes.

2. Review of Spatial Regression Models: Theoretical Implications

2.1. Spatial lag model

The spatial lag model takes the form

\[ y = \rho Wy + X\beta + \epsilon \]
\[ \epsilon \sim N(0, \sigma^2 I_n) \]

where \( y \) is a \( n \times 1 \) vector of house prices, \( W \) is the pre-specified \( n \times n \) matrix of spatial weights which relates the sale price of one house to the sale prices of other houses in the sample by specifying a neighborhood for each house, \( \rho \) is a spatial autocorrelation coefficient (when \( W \) is row-standardized), \( X \) is a \( n \times k \) matrix of explanatory variables including house structural characteristics, location characteristics and environmental attributes of interest, \( \beta \) is a \( k \times 1 \) vector of parameters to be estimated, and \( \epsilon \) is a \( n \times 1 \) vector of errors. The assumption of normality for \( \epsilon \) is necessary for ML estimation.

We can rewrite the lag model as \( y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \epsilon \). Conditioning on \( X \), the covariance structure of \( y \) takes the form of \( \sigma^2 (I - \rho W)^{-1} (I - \rho W)^{-1} \). Since \( (I - \rho W)^{-1} \) is a full matrix with \( \rho \neq 0 \), a first implication is that the house prices are globally correlated, since each location is correlated with every other location in the system, although in a fashion that decays with order of neighbors especially when \( W \) is very sparse. More importantly, it says that house price at one location depends on the characteristics (\( X \)) of all other houses in the sample.
and the errors of all other observations. This outcome can be easily seen when $W$ is row-standardized. Footnote 8 of Kelejian and Prucha (1998) allows for the “Leontief expansion” of

$$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots,$$

then we have

$$y = (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots)X\beta + (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots)\epsilon$$

$$= X\beta + \rho WX\beta + \rho^2 W^2 X\beta + \ldots + \epsilon + \rho W\epsilon + \rho^2 W^2 \epsilon + \ldots \quad (2)$$

The spatial lag specification of house price is justified under the real-estate appraisal process using comparable sale prices or by the herd effect of house price externality. However, Bell and Bockstael (2000) argued that this motivation for spatial lag structure in house prices is weak. It would occur only when the selling price reacts directly to selling prices of neighboring houses and not just to characteristics of neighboring houses. Furthermore, this relationship would only be one-way, namely, the first sold house would affect subsequent sales but not vice versa. The contemporaneous spatial lag structure can be motivated by the externality argument, however, where homeowners enjoy amenity benefits from living near high-priced homes.

One important characteristic of housing data is that it is often measured with error. House transaction databases are often collected for purposes of property tax assessment. This can lead to measurement errors in several ways, even for arms length sales. First, while the recorded sale price will tend to be reliable, the negotiated sale could include items not considered as part of the house, such as an allowance for new flooring, payment of loan or closing costs by the seller, or additional goods that are sold along with the house such as a hot tub. These will result in mismeasurement of the sale price. Further, structural characteristics of the house are often mismeasured. A common example is when a homeowner adds finished square footage to the home (for example by finishing a basement) but does not notify the assessment office. The
recorded square footage of the house at the time of the sale would be less than the actual square footage.

Mismeasurement of either house price or structural characteristics should be captured in the error term, $\varepsilon$. In the spatial lag model, however, that error term will affect all neighboring house prices, and in fact will affect all house prices in the dataset, though the impact will decay with distance. For example, if a house is sold and additional goods such as a hot tub are included in the sale price, the model will expect that the resulting higher sale price will spill over to neighboring houses, raising their prices, which will in turn spill over to the next round of neighbors, and so on. In reality, the higher sale price due to inclusion of the other goods would likely have no impact on neighboring sale prices, except to the extent that real estate agents and appraisers incorrectly view the sale price as a signal of the value of the sold house, and consequently assign higher values to subsequent neighboring houses.

2.2. SAR error model

While the spatial lag model captures the idea that one house can directly influence the value of neighboring houses, spatial error models capture the idea that if the location of a house influences its selling price, nearby houses will be affected by the same location factors. This could happen in two different ways (Durbin 1998). First, since the inclusion of all relevant location attributes is seldom fulfilled and the effects of all omitted variables are subsumed into the error term, if omitted variables are spatially correlated, so are the regression errors. For example, if a local disamenity is not controlled for, it will depress house prices for all nearby houses. Second, measurement error in location variables may exist, which would be spatially
correlated. For example, if air quality is mismeasured for one neighborhood, the error will affect all houses in that neighborhood.

The error term can take different forms of spatial structure. The spatial autoregressive (SAR) process is the most popular. It is similar to the counterpart in a time-series context. The SAR error model takes the form

\begin{align*}
y &= X\beta + \varepsilon \\
\varepsilon &= \lambda W \varepsilon + u \\
u &\sim N(0, \sigma^2 I_n)
\end{align*}

where \(\lambda\) is a spatial parameter similar to \(\rho\) in (1), \(u\) is a \(n \times 1\) vector of i.i.d. errors, and all other notations are as previously defined. In this specification, the error for house \(i\) depends on the average of the errors in neighboring observations, \(\lambda W \varepsilon\), and its idiosyncratic component, \(u_i\).

Solving for \(\varepsilon\) as \(\varepsilon = (I - \lambda W)^{-1} u\), we have \(E(\varepsilon \varepsilon') = \sigma^2 (I - \lambda W)^{-1} (I - \lambda W')^{-1}\), a full matrix with \(\lambda \neq 0\). Thus, every house’s error is correlated with other house’s error in the system, with a distance decay effect. This means that in the SAR model, if square footage is mismeasured for one observation, \(i\), the resulting idiosyncratic error \(u_i\) will spill over to all other observations in the dataset, through the spatial multiplier \((I - \lambda W)^{-1}\).

2.3. Spatial error components (SEC) model

In both the spatial lag model and the SAR error model, any error that affects one house’s price must also affect neighboring house prices. However, measurement error of square footage for one particular house is house-specific. The buyer and the seller know the correct square footage. It is only the researcher (and the tax assessor) who is misinformed. The resulting estimated error
should not spillover to its neighbors. The SEC model, in contrast, allows for idiosyncratic errors that do not spill over to neighboring houses.

The spatial error components model was originally proposed by Kelejian and Robinson (1993, 1995) to avoid the singularity problem associated with the SAR process of the dependent variable/error term. The SEC model combines both a local error term and a spillover error term in the covariance structure for the error term in regression, taking the form

\[ y = X\beta + \varepsilon \]
\[ \varepsilon = W\phi + u \]  

where \( \phi \) is a \( n \times 1 \) vector of errors that spill over across neighbors, \( u \) is a \( n \times 1 \) vector of house-specific errors that do not spill over. Each element in both \( \phi \) and \( u \) is assumed to be iid (or weakly, uncorrelated and id) with mean zero; all other notations are as previously defined. In contrast to the SAR error model, only one component of the composite error terms has a spatial property. This is the spillover error term \( \phi \), which captures unobserved components that are correlated across nearby houses. The location-specific errors \( u \) capture idiosyncratic unobserved components that are specific to each house. This idea is highlighted by the complete error variance-covariance matrix of \( \varepsilon \),

\[ E(\varepsilon \varepsilon') = \sigma_\phi^2 W W' + \sigma_u^2 I \]  

where \( \sigma_\phi^2 \) is the variance component associated with the spillover error and \( \sigma_u^2 \) is the usual local variance term. The only nonzero covariance entries in the variance-covariance matrix are those corresponding nonzero entries in \( WW' \). For first-order \( W, WW' \) consists of first and second order neighbors only. As a result, the SEC model yields zero covariance beyond the second-order neighbor and can be considered as a model for local spatial correlation.
Of the models considered here, the SEC model best captures our intuition about the process that drives spatial correlation in house prices. It explicitly considers the two sources of variation in house price given observed house characteristics. Some errors are “contagious” such that neighboring house prices are observed to be correlated. These errors spill over to neighboring houses via the spatial structure denoted by the spatial weights matrix. Other errors are house-specific in the sense that they will not be transmitted and so have no influence on neighboring houses.

3. Row-standardization of the Spatial Weights Matrix

The spatial weights matrix \( W \) is used to relate an observation at one location to the observations in other spatial units in the system by specifying a neighborhood for each observation. Each element in \( W \) represents whether the two observations are spatially correlated and how strong the relationship is. The selection of a spatial weights matrix is made prior to running the model, in other words, it is not estimated as part of the model.

House transaction data is typically located as points in space (as opposed to polygons). The concept of “contiguity” must therefore be interpreted. Typically, the researcher specifies a maximum distance, and defines any two houses located closer to each other than that distance as contiguous. The choice of distance cutoff is based on past experience or intuition for a given housing market. The \( W \) matrix then contains zeros for all house pairs that are not neighbors, and ones for all pairs that are. An alternative approach is to construct a distance-based weights matrix, where the elements of \( W \) are some decreasing function of the distance between neighbors, to reflect the idea that closer neighbors should be more correlated than more distant neighbors.
Regardless of whether a contiguity-based or a distance-based weights matrix is used, common practice is to row-standardize the weights matrix, so that each row of $W$ sums to unity. This is done by dividing each row by its sum. Using a row standardized $W$, the term $Wy$ will then represent the average price of neighboring houses in the spatial lag model, while $W\varepsilon$ will be the average of the neighboring errors in spatial error models. The associated spatial parameter then has an intuitive interpretation as a spatial autocorrelation coefficient $\gamma$; it also facilitates the maximum likelihood estimation of spatial models. As a result, row-standardization has become a convention in practice. However, row-standardization of spatial weights matrix assumes a specific spatial structure of the dependent variable that may or may not be justified.

3.1. Number effect and distance effect

A common feature of property sales datasets is that the number of neighbors varies across observations, as does the distance to each neighbor. Some houses have many near neighbors while others have few. For some houses, neighbors are located close by, while other houses have closest neighbors located farther away. As the number and density of neighbors for each spatial observation are generally not the same, row-standardization rescales each row of the weights matrix by different factors.

For a given observation, although not changing the relative dependence among all neighbors, row-standardization does change the total impact of neighbors across observations. It is equivalent to assuming that the total effect of neighbors is constant across all observations no matter how many neighbors an observation has. This is what we label the “number effect” of row-standardization. Consider two houses. For one house, data is available for 20 near neighbors.
For the other house, data is available for only 2 neighbors. If the spatial weight is row-standardized, the total effects of neighboring houses are forced to be the same for the two houses.

This may or may not be desirable. In a spatial lag model, unless the weights matrix is row standardized, the lag term $\rho Wy$ will tend to be bigger for observations with more near neighbors. The model will predict that the house with 20 near neighbors will tend to have a higher price than the house with 2, all else equal. That is clearly not appropriate, and a row-standardized spatial weights matrix is clearly preferred for the spatial lag model.

However, the numbers effect could be interpreted differently in the spatial error models. In contrast to the spatial lag model, where every $y$ is positive, and additional observations always increase the value of $\rho Wy$, in the SAR error and SEC models, the errors could be positive or negative. If the weights matrix is contiguity based, then $\lambda W \epsilon$ in the SAR error model and $\lambda W \phi$ in the SEC model represent the average of neighboring errors if $W$ is row standardized, and represents the sum of neighboring errors if $W$ is not row standardized. Both measures have zero expectation. Which measure is preferred? One could argue that, for reasons similar to the argument made above for the spatial lag model, the size of the impact should be the same for each observation, favoring the row standardized matrix. However, an opposing argument exists.

The terms $\lambda W \epsilon$ and $\lambda W \phi$ could be seen as signals of the net impact of unobserved spatial variables. If these terms are large and positive, it implies that house prices tend to be higher than expected in this neighborhood. The amount of information contained in this signal should depend on how many observations go into its construction. A signal based on many neighbors is more informative than one based on few neighbors. Following this argument, $\lambda W \epsilon$ and $\lambda W \phi$ should tend to get larger in magnitude as the number of neighbors increases, favoring use of the non-standardized weights matrix.
The second issue to be considered is what we call the “distance effect”. For a contiguity-based spatial weights matrix, where each neighbor is assigned the same weight of unity, the influence of any one neighbor for a given observation is inversely proportional to the number of total neighbors after row-standardization. While for a distance-based spatial weights matrix, where the interdependence of spatial observations is assumed to decline with distance, this practice rescales the weights based on the absolute distance to neighbors for each row. Again consider two houses. To exclude the number effect, this time suppose that both houses have the same number of neighbors, say 3. For house i, the distances from the three neighbors are 1, 2 and 4. For house j, the distances from the three neighbors are 2, 4 and 8. Assume an inverse-distance weight as \( w_{ij} = 1/(d_{ij})^\theta \) if \( d_{ij} \leq m \) or \( w_{ij} = 0 \) if \( d_{ij} > m \), where \( d_{ij} \) is the distance between points i and j and m is the critical distance value. Setting \( \theta = 1 \) and \( m > 8 \), then the weights for house i are 1, 1/2 and 1/4 and the weights for house j are 1/2, 1/4 and 1/8, respectively. After row-standardization, the weights for both houses are the same: 4/7, 2/7 and 1/7. As a result, the neighbor located 2 units away from the first observation is given the same weight as the neighbor located 4 units away from the second observation (although closer neighbors still get larger weights for the same observation). In this case, Tobler’s first law of geography is violated: remote neighbors are correlated in the same way as closer neighbors and the effect of distance disappears. For a distance-based weights matrix, this effect favors using a non-row-standardized weights matrix.

To summarize, the numbers effect clearly favors use of a row-standardized weights matrix for the spatial lag model. For the SAR error and SEC models, however, the implication of the numbers effect is ambiguous. The distance effect argues against row standardization for the SAR error and SEC models if a distance-based weights matrix is used. Based on these
arguments, in the empirical application in the next section, the spatial lag model will be estimated using a row standardized weights matrix, while both a row standardized and a non-row standardized matrix will be considered for the SAR and SEC models.

The choice of whether to row standardize can be important empirically. Among others, Bell and Bockstael (2000) find far greater sensitivity of parameter estimates to the specification of the spatial weights matrix than the choice of estimation technique. They also point out that this sensitivity and the concern over the change in spatial structure by row-standardization are related each other. Consequently, caution is needed when row-standardizing the spatial weights matrix.

4. Empirical Analysis: Landfill Externality

The empirical study involves estimation of the disamenity associated with proximity to a municipal solid waste landfill. Many of these studies have found the existence of a negative externality from a landfill on nearby house prices, e.g., Hite et al (2001) combined an urban location choice model with hedonic pricing model to incorporating some spatial effects of disamenity in estimating the impact of landfill on nearby property values. Proximity to a landfill is an example of a class of environmental goods (or bads) that present unique issues in hedonic pricing models because the variation in the level of the good is strongly spatially correlated. In a typical hedonic study estimating the impact of a single landfill on nearby property values, observations (houses) in a dataset that have high levels of the good must be located near each other. This is in contrast to environmental quality measures such as noise or air quality, where the study area can include multiple subareas with high or low values. If a landfill happens to be located in an area that has higher or lower than usual house prices due solely to random effects,
and those random effects are correlated across nearby observations, OLS estimation of the hedonic price function can lead to erroneous conclusions about the sign and statistical significance of the marginal implicit price of landfill proximity.

The purpose of the empirical study is to estimate the MIP of landfill proximity using OLS and the three spatial models described in Section 3. The performance of the spatial models will be compared using criteria based both on goodness of fit and Lagrange Multiplier statistics, and a preferred model will be chosen. The model estimations will then be compared to determine how robust the estimated MIP’s for landfill proximity are across models.

4.1. Data and variable description

The data set under study consists of 7493 arm-length house sales between 1998 and 2002 in southern and eastern Berks County, Pennsylvania. The study area includes three landfills. Pioneer Crossing Landfill (PCL) is prominent on the landscape and directly visible from many nearby residential neighborhoods. Rolling Hills Landfill (RHL) is the largest of the three and is visible from at least some vantage points, although surrounding topology shields it from view from most directions. Western Berks Landfill (WBL) is the smallest and is physically isolated from residential areas and difficult to see from off property.

Houses included in the dataset are located within 10 kilometers of one of the three landfills. Data on sale price, sale year, township, location and structural characteristics of the house are obtained. An initial analysis (Ready 2005) used dummy variables to define concentric rings around each landfill. In that analysis, the coefficient for the dummy variables for houses located within two miles of PCL were negative and statistically significant, but those for houses located outside of two miles were not. For RHL, the coefficients were negative and statistically
significant for houses within three miles, but insignificant for houses outside three miles. For WBL, none of the dummy variables were statistically significant.

Based on these results, an index of proximity to each landfill was constructed of the form

\[
\text{DI}(i) = \begin{cases} 
  d_i & \text{if } d_i < \bar{d}_i \\
  \bar{d}_i & \text{otherwise}
\end{cases}
\]

where \(d_i\) is the distance (in miles) from the house to landfill \(i = \text{PCL, RHL, and WBL}\), and \(\bar{d}_i\) is the outer limit of the landfill’s impact on house prices, set equal to 2 miles for PCL and 3 miles for RHL. Even though the initial analysis suggested that WBL did not have an impact on nearby property values, a proximity index was constructed for it as well with an outer limit of 2 miles, to see whether any of the estimated models found a statistically significant MIP for that landfill.

One complicating factor when estimating a MIP for landfill proximity is that landfills are an industrial land use, and are often surrounded by other land in industrial use. If industrial land has a negative impact on nearby house prices (as found, for example, by Ready and Abdalla) then failing to control for the presence of industrial land will lead to omitted variable bias in the MIP for landfill proximity. For this reason, two measures of surrounding land use were included in the hedonic price models. These were the proportion of land within 400 meters of the house in industrial use, and the proportion of land between 400 and 800 meters from the house in industrial use. Preliminary analysis showed that industrial land located farther than 800 meters from the house did not impact house price.

Data on structural characteristics of the house (square footage, age, physical condition, etc.) and on the most recent sale (price and date) were obtained from the Berks County Office of the Assessment in GIS format. The dependent variable in the hedonic regression is the natural
log of real sale price, measured in 2002 dollars. Other house characteristics included in the regression included commuting distance to Philadelphia, and Allentown, elevation and slope at the house site, and a measure of school district quality. For details, see Ready (2005).

4.2. Choice of spatial weights matrix

For each residential parcel sold in the data set, the location of the house is determined as the centroid of the residential parcel. A spatial weights matrix is constructed based on the distances between these centroids. The choice of boundary for the definition of a neighborhood is not clear-cut. Ready and Abadalla (2005), in the same county, found that nearby land use can impact house prices up to 1600 meters away. This provides one possible definition of a “neighborhood.” We also use 1600 meters as the cutoff distance for the spatial weights matrix, but recognize that this choice is somewhat ad hoc.

Within 1600 meters, we believe that spatial correlation will be stronger for houses located closer to each other. We therefore choose spatial weights that vary with distance 5:

\[
W_{ij} = \begin{cases} 
1 - \frac{d_{ij}}{L} & d_{ij} < L \\
0 & d_{ij} \geq L \text{ or } i = j 
\end{cases}
\]

where \(d_{ij}\) is the distance between house \(i\) and house \(j\), and \(L\) is the cutoff distance (1600 meter).

4.3. Model selection

In Section 2, we argue that the SEC model is better suited to house sales data than the SAR error or spatial lag models. There also exist two empirical methods 6 for choosing among these competing models, LM tests and a goodness of fit criterion. Neither of these empirical methods
is decisive. Each may incur some sources of misspecifications. It is best to consider both methods to see if they provide consistent guidance.

4.3.1 LM tests

Test statistics based on the Lagrange Multiplier (LM) principle have been developed for detecting the existence of spatial correlation in the regression error and of a spatially lagged dependent variable (Anselin 1988, Anselin et al 1996, Anselin and Bera 1998, Anselin 2001). There are two main types of LM tests, one-directional tests and robust tests. One-directional tests are designed to test a single specification by assuming that the rest of the model is correctly specified. The test statistics are developed for the null hypothesis of $H_0: \rho = 0$ assuming $\lambda = 0$ for the spatial lag model or $H_0: \lambda = 0$ assuming $\rho = 0$ for the SAR error model. Although it is not possible to develop robust tests in the presence of global misspecification ($\lambda$ and $\rho$ take values far from zero), Anselin et al (1996) propose LM tests for the spatial lag model and SAR error model which are robust to local misspecification. Furthermore, Anselin and Moreno (2003) develop a LM test for the SEC model. The LM-SEC statistic performs well in terms of power against the null hypothesis of no spatial correlation in Monte Carlo experiments, even in non-normality situations. There is no such test to distinguish the SEC model from the spatial lag or the SAR error model, and no robust LM test for the SEC model.

The procedure of using LM test statistics for the spatial lag model and the SAR error model follows the decision rule of Anselin (2005). This decision rule has been adopted by other empirical studies, e.g., Kim et al, 2003. Based on the OLS regression residuals, LM-Lag test statistics for a spatially lagged dependent variable and for SAR errors were calculated using both row-standardized and non-standardized spatial weights matrices. All four tests show highly
significant results and lead to rejection of the null hypothesis of no spatial correlation in each case (Table 1). The LM-SEC test for the SEC model is also highly significant for both weights matrices. For completeness, the Moran’s I test is also included.

Since both LM tests for the spatial lag and SAR error structures reject the null hypothesis, we need to consider their robust forms. Before proceeding to next step, a brief look at Table 1 shows LM-Lag ($W_s$) > LM-Lag ($W_n$) and LM-Error ($W_n$) > LM-Error ($W_s$). The lag structure with row-standardized weights matrix is more significant while the error structure with non-standardized weights matrix is more significant. These results are consistent with the claims made in section 3 that a row-standardized weights matrix is more appropriate for the spatial lag model and a non-standardized weights matrix may be more suitable for the SAR error model. We also have LM-SEC ($W_s$) > LM-SEC ($W_n$), suggesting that, a row-standardized weights matrix appears to better represent the SEC structure. Further, when the SEC model was estimated using a non-standardized weights matrix, the estimated spillover error variance for the SEC model, $\sigma_{\phi}^2$, was negative, implying that the SEC model with the non-standardized weights matrix does not fit to our dataset (see footnote 12 of Anselin and Moreno, 2003).

The robust LM test involves two weights matrices, one for the lag structure and the other for the SAR error structure. There exist four combinations of the two matrices. Referring to Table 2, all the eight robust test statistics are highly significant (some are much greater than the non-robust tests). Furthermore, we find Robust LM-Lag ($W_s$,$W_n$) > Robust LM-Lag ($W_n$,$W_n$) and Robust LM-Lag ($W_s$,$W_s$) > Robust LM-Lag ($W_n$,$W_s$), which imply that, in the presence of possible locally misspecified SAR error structure, the lag structure with a row-standardized weights matrix is more significant than that with a non–standardized weights matrix. We also
find that Robust LM-Error ($W_n, W_n$) $>$ Robust LM-Error ($W_n, W_f$) and Robust LM-Error ($W_f, W_n$) $>$ Robust LM-Error ($W_f, W_f$), which imply that, in the presence of a possible locally misspecified lag structure, the SAR error structure with a non-standardized weights matrix is more significant than that with a row–standardized weights matrix. As in the non-robust tests, these robust LM tests support the previous claims that a row–standardized weights matrix better captures the lag structure, while a non-standardized weights matrix better captures the SAR error structure for this dataset.

Next we compare the Robust LM-Lag ($W_f, W_n$) with the Robust LM-Error ($W_f, W_n$), with the favored weights matrices for each of the spatial structures. The latter (1556.50) is much greater than the former (166.56), suggesting that among all possible SAR and spatial lag models, the SAR model with the non-standardized weights matrix best captures the spatial structure of this dataset. A similar comparison between the SAR model and the SEC model is not possible. Those two competing models must be compared using a different criterion.

4.3.2. Pseudo-$R^2$ criterion

We compare the SAR error and SEC models using a goodness-of-fit criterion we call the pseudo-$R^2$. For each spatial model, a pseudo-$R^2$ value is calculated equal to 1 minus the ratio of the variance of prediction errors over the variance of the dependent variable. A model with higher pseudo-$R^2$ value is preferred. For the OLS model, this is the standard $R^2$ statistic. For the spatial models, this statistics is called “pseudo” $R^2$ because the sum of the predicted errors is not guaranteed to be zero. However, if the average predicted price is close enough to the average actual price, as it is in our case, the intuition behind the pseudo-$R^2$ is similar to that of the usual $R^2$ statistic.
In spatial models, a prediction for an individual house sale price could be made based on different information sets. The simplest information set would include only the estimated model parameters ($\beta$, $\rho$, and/or $\lambda$), and the characteristics of the house ($X_i$). This would be enough to predict a house’s sale price in the OLS, SAR error and SEC error models. To predict a house’s sale price in the spatial lag model, you would also need to know, at least, the characteristics of all other houses, i.e. the full $X$ matrix. Because the spatial lag model uses more information to make predictions than does the OLS, SAR error or SEC models, the pseudo-$R^2$ is not a good criterion for comparing the spatial lag model to the other three models. However, it can be used to compare among the OLS, SAR error and SEC error models. For a more detailed discussion of the use of the pseudo-$R^2$ criterion within each spatial model, see Wang (2006).

The calculated pseudo-$R^2$ statistics were 0.843 for the SEC ($W_s$) model, 0.812 for the SAR ($W_s$) model, and 0.815 for the SAR ($W_n$) model. We conclude that the SEC($W_s$) model is the most appropriate model for our dataset.

4.4. Model estimation

The preceding section shows that the SEC model with row-standardized spatial weights matrix is the preferred model. However, it is of interest to explore whether the estimated MIP values are sensitive to model choice. Hedonic regression models were estimated for the OLS, Lag($W_s$), SAR($W_a$), SAR($W_s$), and SEC($W_s$) models. To estimate the lag model, we adapt the IV/2SLS estimation, in which $WX$ acts as an instrument for $Wy$ to avoid the introduced endogeneity problem. For this dataset, the ML estimates for the row-standardized lag model are very close to the IV/2SLS estimates. In addition, compared with the ML estimation, the IV/2SLS method allows for a 2SLS-robust estimation by correcting possible heteroskedasticity. For the SAR error
model, we use the generalized moments (GM) method suggested by Kelejian and Prucha (1999) to estimate the spatial parameter ($\lambda$), followed by a Cochran-Orcutt-type transformation to account for the estimated error structure for the SAR error model. For the SEC model, a general method of moments (GMM) estimator (Anselin and Moreno, 2003) for the two variance components is first obtained, followed by the same feasible GLS procedure as the SAR error model through Cholesky factorization.

Complete regression results for all models are available from the authors. Coefficient estimates for house structural characteristics (square footage, age, condition, etc.) and location characteristics (commuting distances, school district quality) are consistent with expectations, and in all cases statistically significant. The impact on property prices of industrial land within 400 meters of the house was negative and statistically significant for all models. The impact of industrial land between 400 and 800 meters away from the house was negative for all models, but was statistically significant only for the OLS model.

Table 3 shows the MIP estimates for all models for the three landfill proximity indices. The preferred model, the $SEC(W)$, is shown in the far right column. With that model it is found that property values are depressed within 2 miles and 3 miles of two prominent and visible landfills, but not affected by the less-prominent landfill (WBL).

Taking the $SEC(W)$ as the preferred set of MIP estimates, we can determine the sensitivity of the MIP estimates to model choice. Note that for the spatial lag model, the MIPs of house attribute $i$ for all houses are $\beta_i (I - \rho W)^{-1}$, which is different from the estimate ($\beta_i$) of a spatial error model or traditional OLS hedonic model (see Kim et al 2003). In the spatial lag model, the total impact of a landfill on nearby house prices includes the direct impact (captured by $\beta_i$) plus the indirect impact through the landfill’s direct impact on neighboring house values.
(plus higher order impacts). The true MIP of landfill proximity for the spatial lag model will therefore be larger than the parameter estimates in column 2 of Table 3.

With some exceptions, the various models give MIP estimates that are similar in sign, size, and statistical significance. The parameter estimates are particularly consistent for RHL, where all parameter estimates are positive and statistically significant, indicating that proximity to RHL has a negative impact on house prices, and the parameter estimates are all close in size to the estimated parameter for the preferred SEC($W_s$) model. As expected from the discussion above, the spatial lag model has the smallest parameter estimate. For PCL, the estimated parameters are positive for all models, and smaller in size than for RHL. The estimated parameters are all statistically significant except for the SAR($W_s$) model, which has a parameter estimate smaller than for other models, and statistically insignificant. In all models except the $	ext{Lag}(W_s)$ model, WBL is found to have no significant impact on nearby house prices, consistent with the preliminary analysis. In the $	ext{Lag}(W_s)$ model, WBL is found to actually have a positive impact on nearby prices, a result that we cannot explain, and one that suggests possible model mis-specification.

5. Discussion

All evidence shows that there is strong spatial structure in this house price dataset, suggesting that OLS will result in biased and/or inefficient parameter estimates, and biased standard error estimates. Still, the estimated MIP’s for landfill proximity were surprisingly robust across alternative models.

Interestingly, OLS yielded parameter estimates that were very similar to the preferred spatial model. This was true for all landfill MIP’s and for the measures of nearby industrial land.
However, the OLS model had larger t-statistics for almost all of these variables than did the preferred SEC($W_s$), consistent with the idea that if the data has spatially correlated errors, OLS will lead to consistent parameter estimates, but biased standard error estimates.

In two cases, models generated parameter estimates that would lead to different conclusions than those from the preferred model. The SAR ($W_s$) model gave an estimated MIP for PCL that was much lower than that from the SEC($W_s$), and was not statistically significant. The Lag($W_s$) model gave an estimated MIP for WBL that was statistically significant and had sign opposite to expectations.

For this dataset, then, we find that OLS gave results that were closer to the preferred model results than did two commonly used spatial models. While this pattern of results is specific to our dataset, it does point out the danger in uncritically applying spatial models to hedonic house price datasets. Using an inappropriate spatial econometric model may be worse than using OLS, even in situations where there is strong spatial structure in the data.

To summarize, this paper argues that the SEC model is better suited on theoretical grounds to use in hedonic house price analyses than is either the spatial lag model or the SAR error model. The SEC model is preferred because it best represents the intuition about the process that drives the spatial correlation of house prices by explicitly specifying two sources of variations in house price given house characteristics. We find that for our dataset, the SEC model is the preferred model, according to LM and goodness of fits tests. Finally, we find that parameter estimates are largely robust to the choice of spatial models, but that use of an inappropriate spatial model can lead to incorrect inferences.

We also consider the issue of row standardization of the spatial weights matrix in spatial models. For house price data, we argue that row standardization is unambiguously desirable in
the context of the spatial lag model, but that we cannot on theoretical grounds determine whether
the weights matrix should be row standardized in the spatial error models. Empirical tests
suggested that, for our dataset, row standardization better modeled the spatial structure within the
SEC model, but that in the SAR model, the data was better modeled by a non standardized
matrix.

We encourage those conducting hedonic pricing studies to explore alternative spatial
models, and in particular we encourage them to consider the SEC model. The choice of which
model is preferred will likely vary from dataset to dataset, but should be based on a combination
of theoretical arguments and empirical evidence.
Notes:

1. There are some differences among authors in how these models are labeled. The model that is here called the spatial lag model is called by LeSage (1999) the spatial autoregressive (SAR) model, while the model called here the SAR error model is called by LeSage the spatial error model (SEM).

2. Another popular specification is spatial moving average (SMA) process. The model with SMA error sees rare applications although a corresponding LM test has been developed. One reason is that it introduces extra difficulties in ML estimation (Sneek and Rietveld, 1997) and no other widely accepted estimation method is available.

3. The singularity problem refers to the possibility of zero determinant for $|I - \rho W|$ in the spatial lag model or $|I - \lambda W|$ in the SAR error model such that $I - \rho W$ or $I - \lambda W$ is not invertible.

4. See Kelejian and Robinson (1995, 76-77) for discussion about this point.

5. This weight differs from the linear proximity index just by a factor $d_{ij}$. Another specification of bi-square weights produces similar estimation and test results for all spatial models considered. As has been shown in other studies (e.g., Bell and Bockstael, 2000), the estimation outcomes are not very sensitive to the specification of weight function form.

6. A third selection method is the Bayes factor method. Hepple (2004) developed the formulae for calculating Bayes factors for a family of spatial regression models including the spatial lag
model, the SAR error model, the SMA error model, etc. However, in large sample cases, the Bayes factor method is subject to numerical problems (see Wang 2006).

7. LeSage’s Spatial Econometrics Toolbox includes the routines of LM-Lag, LM-Error and Moran’s I tests for row-standardized weights matrix. However, these routines are not feasible (or are at least computationally inefficient) for large-sample problem, as in our case. We develop Matlab routines for all LM tests used in this section, which are feasible and efficient for large-sample problem. These routines are available from the authors upon request.
References


**Table 1  LM and Moran’s I tests**

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>Spatial Weights Matrices</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_s$</td>
<td>$W_n$</td>
</tr>
<tr>
<td>LM-Lag</td>
<td>314.37</td>
<td>27.00</td>
</tr>
<tr>
<td>LM-SAR</td>
<td>825.08</td>
<td>1704.30</td>
</tr>
<tr>
<td>LM-SEC</td>
<td>734.49</td>
<td>452.50</td>
</tr>
<tr>
<td>Moran’s I</td>
<td>37.48</td>
<td>72.47</td>
</tr>
</tbody>
</table>

**Table 2  Robust LM tests for the spatial lag model and the SAR error model**

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>Spatial Weights Matrix Combination$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>($W_s, W_s$)</td>
</tr>
<tr>
<td>Robust LM-Lag</td>
<td>133.64</td>
</tr>
<tr>
<td>Robust LM-Error</td>
<td>644.36</td>
</tr>
</tbody>
</table>

Note: $^a$. The first spatial weights matrix in parenthese is for lag structure and the second one for SAR error structure.
Table 3  Estimation results of the OLS and spatial regression models

<table>
<thead>
<tr>
<th>Parameters</th>
<th>OLS</th>
<th>Lag($W_s$)</th>
<th>SAR($W_s$)</th>
<th>SAR($W_n$)</th>
<th>SEC($W_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DI (PCL)</td>
<td>0.0983</td>
<td>0.0436</td>
<td>0.0257</td>
<td>0.0682</td>
<td>0.0674</td>
</tr>
<tr>
<td></td>
<td>(6.85)</td>
<td>(3.03)</td>
<td>(0.96)</td>
<td>(3.84)</td>
<td>(3.33)</td>
</tr>
<tr>
<td>DI (WBL)</td>
<td>-0.0156</td>
<td>-0.0459</td>
<td>-0.0076</td>
<td>-0.0011</td>
<td>-0.0254</td>
</tr>
<tr>
<td></td>
<td>(-1.03)</td>
<td>(-3.08)</td>
<td>(-0.32)</td>
<td>(-0.06)</td>
<td>(-1.28)</td>
</tr>
<tr>
<td>DI (RHL)</td>
<td>0.0716</td>
<td>0.0618</td>
<td>0.0908</td>
<td>0.0715</td>
<td>0.0807</td>
</tr>
<tr>
<td></td>
<td>(3.94)</td>
<td>(3.48)</td>
<td>(2.66)</td>
<td>(3.96)</td>
<td>(3.08)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.2230</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(16.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda^a$</td>
<td></td>
<td>0.6803</td>
<td>0.0040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2\rho$</td>
<td></td>
<td></td>
<td></td>
<td>0.0639</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(6.51)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_u$</td>
<td></td>
<td></td>
<td></td>
<td>0.0212</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(44.01)</td>
<td></td>
</tr>
<tr>
<td>Jarque-Bera $^b$</td>
<td>476.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: t-statistics in parentheses

$^a$ The GM estimation for the SAR error model does not produce a variance estimate for the spatial parameter;

$^b$ Jarque-Bera test against the null hypothesis of error normality is rejected with high probability.