JOURNAL OF REGIONAL SCIENCE, VOL. 50, NO. 1, 2010, pp. 119–141

ISSUES IN SPATIAL DATA ANALYSIS

Daniel P. McMillen
Department of Economics and Institute of Government and Public Affairs, University of Illinois (MC-037), 1007 W. Nevada St., Urbana, IL 61801.
E-mail: mcmillen@illinois.edu

ABSTRACT. Misspecified functional forms tend to produce biased estimates and spatially correlated errors. Imposing less structure than standard spatial lag models while being more amenable to large datasets, nonparametric and semiparametric methods offer significant advantages for spatial modeling. Fixed effect estimators have significant advantages when spatial effects are constant within well-defined zones, but their flexibility can produce variable, inefficient estimates while failing to account adequately for smooth spatial trends. Though estimators that are designed to measure treatment effects can potentially control for unobserved variables while eliminating the need to specify a functional form, they may be biased if the variables are not constant within discrete zones.

1. INTRODUCTION

The goal of many empirical studies in urban economics, regional science, and geography is to measure the effects of proximity. For example, in its simplest form the Alonso–Muth–Mills model of urban spatial structure predicts that distance from the city center is the sole determinant of the spatial variation in variables such as land values, population density, and the per-unit price of housing. In this simple single-explanatory variable model, a primary econometric issue is functional form: while theory predicts these variables will decline smoothly with distance from the city center, the exact shape of the function is ultimately an empirical issue. Other studies attempt to isolate the effects of proximity to a site while controlling for the effects of other variables. For example, a plethora of hedonic housing studies attempt to determine how sales prices vary by proximity to airports, highway interchanges, toxic waste sites, etc. While the results of such studies can be highly sensitive to functional form assumptions, multiple explanatory variable models are made more complicated by the necessity of controlling for the effects of other variables that may be highly correlated with the one of central interest. Is it the presence of toxic waste that is leading to lower house prices, or is it the fact that the area has poor access to areas with growing employment and is filled with badly

Received: April 2009; revised: August 2009; accepted: September 2009.

 DOI: 10.1111/j.1467-9787.2009.00656.x

maintained, outdated housing? Unless the study includes controls for all variables that influence house prices—and they are measured properly, and the functional form is correct, and so on—correctly measuring the effect of any single variable is an extraordinarily difficult task.

In addition to functional form issues, a critical issue in spatial data analysis is that important variables are highly correlated and no study includes all relevant variables. Consider the superficially simple task of testing the monocentric city model's prediction that the price of a unit of housing declines with distance from the city center. It is easy to assemble a large dataset with sales prices and distances. But the model predicts that it is the price of a single *unit* of housing that declines with distance. Since housing does not have well-defined units, it becomes necessary to control for the quantity of housing by using variables like square footage, the number of bedrooms, the presence of a garage, lot size, and so on. Some variable is always missing no matter how long the variable list gets. A missing measure of quality, for example, will tend to be correlated with income. If higher-income households tend to live farther from the city center, it should not be surprising to find that house prices *rise* with distance from the city center even when the model predicts the opposite and even if the per-unit price of housing does in fact decline with distance.

The problem of spatially correlated missing variables is endemic and is not confined to direct measures of proximity. For example, one of the issues examined in the empirical section of this paper is the effect of violent crime on house prices. The measure of violent crime—the number of homicides committed in a census tract during a year—is highly concentrated spatially. Since murders are more common in low-income districts, house prices will tend to be negatively correlated with the homicide rate. Thus, the model has a spatial dimension even though the analysis is not explicitly spatial at first glance.

The twin issues of functional form and spatially correlated missing variables have largely been treated separately in the empirical literature. Functional form choice is typically addressed directly using series expansions or nonparametric estimation procedures. Spatially correlated missing variables are often considered only indirectly through tests for spatial autocorrelation. The standard models used to address spatial autocorrelation are based on ad hoc specifications of a spatial weight matrix. A common approach is to begin with a simple functional form, test for spatial autocorrelation, and then to estimate a model that includes a spatially lagged dependent variable or that accounts directly for spatial autocorrelation in the error terms. After perhaps including some experimentation with different specifications of the spatial weight matrix, further specification testing typically stops. The main problem with this approach is that it is likely to fail in identifying the root cause of the spatial autocorrelation. Functional form misspecification can itself cause residuals to be spatially correlated. More importantly, if the underlying problem is that a spatially correlated variable has been omitted from the regression, a misspecified spatial econometric approach may be accepted in place of the true model.

Unfortunately, the real solution—adding the omitted variable to the analysis—may not be feasible if the data are not available. In this case, the role of further statistical testing is to assess the robustness of the results to alternative model specifications. We can have more confidence in the results if a variety of model specifications lead to similar results. Note, however, that this view of standard spatial econometric models as just another specification test is directly at odds with the classical approach under which they typically have been derived. Standard spatial econometric models are often estimated by maximum likelihood methods. Maximum likelihood provides consistent and efficient estimates if the full model—including the functional form and the error distribution—is known in advance. But in many cases the whole reason for estimating the spatial econometric version of the model is that the base equation produced spatially autocorrelated errors. In addition, standard models require the manipulation of large matrices, impeding their use for large samples. The paradox of most spatial econometric models is that they are limited to small-to-medium samples and their very use is an admission that the true model structure is unknown, yet maximum likelihood estimators rest on an assumption that the true structure is known and can require large samples to produce accurate results.

The focus of econometric modeling shifts once one accepts that obtaining consistent, efficient estimates of a known model structure is a virtual impossibility in spatial models. Standard spatial econometric models become just another tool to guide the ultimate model specification and to assess the robustness of the results. Nonparametric and semiparametric models are attractive alternatives to parametric alternatives because they admit at the start that the true model structure is unknown. Unfortunately, the voluminous statistical literature on nonparametric and semiparametric models has made only limited inroads into standard spatial econometric practice. The most commonly used procedure, bearing the seemingly innocuous but ultimately pernicious sobriquet of "geographically weighted regression," is a special case of standard nonparametric regression procedures. By focusing on this special case, the advantages of other estimators have been neglected. In addition, many researchers fail to understand that nonparametric procedures are not necessarily profligate users of degrees of freedom, that they can be used to conduct hypothesis tests, that they can be implemented easily, and that they can provide reliable estimates of both the predicted values of the dependent variable and the marginal effects of the explanatory variables.

The issue of model specification may be particularly important when the objective of a study is to evaluate the effects of policy decisions. For example, hedonic price function estimates might be used to analyze the effect of zoning decisions on property values or to calculate the benefits of improvements in school quality within district. Model misspecification will directly influence any cost–benefit calculations derived from the hedonic price function estimates. Fortunately, it may be possible to take advantage of estimators designed to measure average treatment effects to obtain more accurate measures of the causal

effects of the policy on property values. These estimators are particularly useful when a policy covers a well-defined, discrete zone such as a school district, an area zoned for a particular land use, or a special tax district. Comparisons of outcomes on either side of a boundary can produce accurate measures of the effects of school quality, zoning, or taxes on house prices or other variables of interest. Caution is still warranted when applying these approaches to spatial data because omitted spatially correlated variables can still bias the results.

The rest of the paper is organized as follows. In Section 2, I use a simple single-explanatory variable to illustrate some of the advantages of series expansion and nonparametric approaches for analyzing spatial data. I also compare the results to a standard spatial econometric model, which in this example is simply another method of generalizing a simple functional form. Next, Section 3 discusses the use of nonparametric approaches for analyzing multivariate models. In Section 4, I discuss situations in which fixed effects estimators may or may be preferable to a semiparametric or nonparametric approach. Section 5 discusses some of the issues encountered with spatial data when using estimators that are designed to measure treatment effects. Section 6 offers some conclusions.

2. SINGLE-EXPLANATORY VARIABLE MODEL

A simple single-explanatory variable model serves as a good introduction to commonly used procedures for analyzing spatial data. As developed by Alonso (1964), Mills (1972), and Muth (1969), the monocentric model continues to serve as the base for most empirical studies of urban spatial structure. Simple versions of the model predict that such variables as land values, population density, and the per-unit price of housing will decline monotonically with distance from the city center. Examples of empirical tests of these predictions Ahlfeldt and Wendland (2008), Anderson (1982, 1985), Atack and Margo (1998), Coulson (1991), McMillen (1996), and Mills (1969).

Somewhat surprisingly, the only paper testing one of the more important predictions of the model—that capital/land ratios decline with distance from the central business district (CBD)—is McMillen (2006). In that paper, I used the floor area ratio (i.e., building area divided by land area) as a measure of capital intensity. I use a similar dataset here to illustrate some basic tools for spatial data analysis. As in the previous paper, the dataset comprises small-scale (six units or less) residential properties in Cook County, an area of over 5 million people, including Chicago and many of its suburbs. Using 2003 assessment data for 1,006,047 properties, I calculate average building areas and average land area across 1,322 census tracts.[1] The primary explanatory variable for

[1]Building and land areas are both measured in square feet. The dataset in McMillen (2006) differs slightly in that it was drawn from 1997 assessment data rather than 2003. More significantly, the base dataset for the earlier paper was smaller since I restricted the analysis to properties that sold between 1983 and 1999 rather than including all assessments.

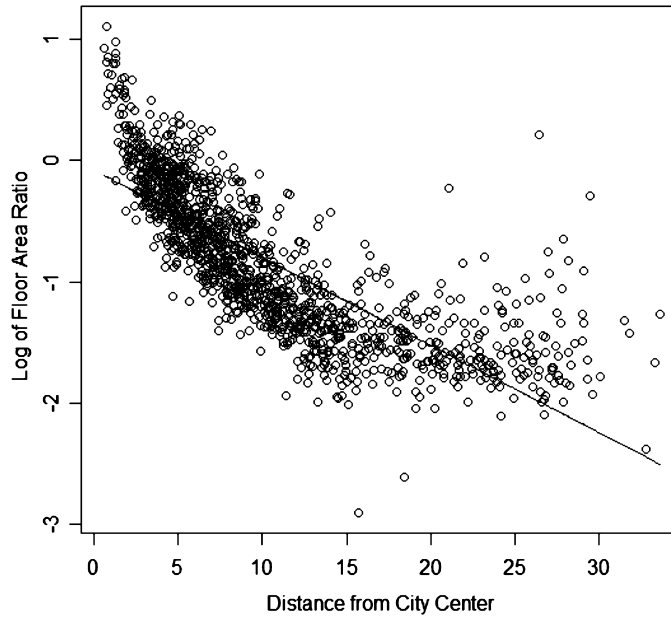

FIGURE 1: Data Plot for Log of Average Floor Area Ratios in Cook County Census Tracts.

the floor area ratio (FAR) is distance from the CBD (x), measured as straight-line miles between the census tract centroid and the traditional city center of Chicago at the intersection of State and Madison Streets.

Figure 1 shows a plot of the raw data. For the first 25 miles, the graph clearly shows a close relationship between the natural logarithm of the floor area ratio and distance from the CBD. Although the function is not linear, the commonly used negative exponential function is clearly a good starting point for the analysis. The first column of results in Table 1 presents the estimates from a regression of $y = \ln(\text{FAR})$ on x. As expected, this simple linear model fits the data well, with distance alone accounting for 61.8 percent of the spatial variation in ln(FAR). The estimated coefficient implies that floor area ratios decline by 7.2 percent with each mile from the city center.[2] The predicted values from the regression form the straight line shown in Figure 1. The true

[2]It is worth repeating here two basic econometric points that are forgotten routinely in this literature. First, in a single-explanatory variable model, R^2's are smaller when the absolute value of the coefficient is smaller, other things being equal. Thus, since the theory predicts that gradients will be smaller as transportation cost declines and (most likely) as income increases, it also predicts that the R^2 will decline unless the variance of the errors declines or the variance of x increases. Second, aggregated data tend to produce higher R^2's than the underlying micro data. The impressive fits evident in Clark's classic (1951) paper remain impressive using current data if one aggregates the data to mile-wide rings around the city center.

TABLE 1: Regressions for the Log of Floor Area Ratios

Variable	Base Linear Model	Spline	Spatial Lag with Linear Base	Spatial Lag with Spline Base
Constant	−0.078	0.626	−0.007	0.172
	(4.017)	(12.634)	(0.572)	(4.220)
x = Distance to City	−0.072	−0.223	−0.013	−0.058
Center	(46.187)	(12.987)	(7.781)	(4.098)
x^2		0.006		7.63*e*-04
		(2.395)		(0.594)
x^3		−7.06*e*-07		2.88*e*-05
		(0.015)		(0.793)
$(x - 16.811)^3$ if $x >$		−2.38*e*-04		−1.65*e*-04
16.811, 0 otherwise		(2.144)		(1.923)
WY			0.833	0.709
			(45.140)	(27.459)
R^2	0.618	0.780		

Notes: Absolute t-values are in parentheses below the estimated coefficients. The number of observations is 1,322.

relationship is clearly nonlinear, with a steeper gradient near the CBD and a flatter gradient at more distant locations.

Series Expansions

The simplest way to account for this nonlinearity is to add powers of x to the estimating equation—x^2, x^3, x^4, etc. Various types of series expansions can often provide better fits with fewer additional variables. Two good choices include the cubic spline (Suits, Mason, and Chan, 1978) and the flexible Fourier form (Gallant, 1981, 1982). The cubic spline involves dividing the axis into S equal intervals ranging from $x_0 = \min(x)$ to $x_S = \max(x)$. The "knots" are the endpoints for the intermediate intervals: $x_1 = x_0 + (x_S - x_0)/S$, $x_2 = x_0 + 2(x_S - x_0)/S, \ldots, x_{S-1} = x_0 + (S - 1)(x_S - x_0)/S$. Associated with each knot is a dummy variable D_s indicating whether x is greater than x_s. The estimating equation is

$$(1) \qquad y = \beta_0 + \beta_1 (x - x_0) + \beta_2 (x - x_0)^2 + \beta_3 (x - x_0)^3 + \sum_{s=1}^{S} \delta_s (x - x_s)^3 D_s + u.$$

Equation (1) shows that the spline simply adds a set of interaction terms between dummy variables and cubic terms to a standard cubic function. The Fourier approach is similar in spirit but starts with a transformation of the explanatory variable, $z = 2\pi(x - \min(x))/(\max(x) - \min(x))$. The Fourier expansion adds trigonometric terms to a base quadratic function:

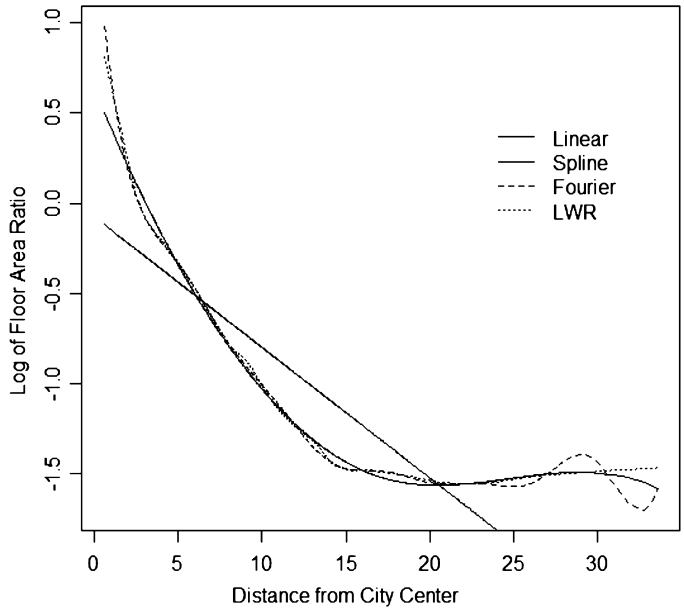

FIGURE 2: Estimated Linear, Spline, Fourier, and LWR Functions.

$$y = \beta_0 + \beta_1 z + \beta_2 z^2 + \sum_{j=1}^{J} \left(\gamma_j \sin(jz) + \lambda_j \cos(jz)\right) + u. \tag{2}$$

For example, if $J = 2$, the explanatory variables are z, z^2, $\sin(z)$, $\cos(z)$, $\sin(2z)$, and $\cos(2z)$. Examples of empirical applications of the spline approach include Anderson (1982, 1985). Applications of the Fourier approach include Ihlanfeldt (2004), McMillen and Dombrow (2001), Thorsnes, Alexander, and McLennan (2009), and Thorsnes and Reifel (2007).

The predicted values for spline and Fourier estimates of the FAR model are shown in Figure 2. The R^2's indicate that at least 78 percent of the variance in ln(FAR) is explained by simple functions of distance.[3] Both the spline and Fourier estimates indicate much steeper gradients near the city center and flatter functions at more distant locations. Consistent with the raw data plot, the Fourier function estimates are steeper than the spline function near the city center. However, the local rise and fall in the function at the distant edge of the dataset, which is driven by a few large values of FAR, appears to be largely spurious and suggests that a lower value of J may be preferable. Even

[3]Although simple visual inspection of the raw data and predictions can be sufficient to determine the order of the expansions for a simple single-explanatory variable model (i.e., S for the spline function and J for the Fourier expansion), in this case I use an Akaike information criterion. The number of expansion terms is $S = 1$ and $J = 5$ using this criterion.

with this local rise in FAR, Figure 2 presents strong support for the monocentric city model's prediction that the floor-area ratio declines smoothly with distance from the city center.

Nonparametric Approaches

Like the Fourier and spline approaches, nonparametric models take the general form $y_i = f(x_i) + u_i$. They impose little structure on the model, instead using simple moving averages or local curve fitting to approximate the function at predefined points.[4] Perhaps the most commonly used nonparametric approach across all fields is kernel regression. Letting x be any arbitrary value of the explanatory variable, the predicted value of y at x is simply a weighted average of the values of y, with weights defined by the kernel function $K((x_i - x)/h)$ and the bandwidth or window size, h. The predicted value is

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} K(\psi_i) y_i}{\sum_{i=1}^{n} K(\psi_i)}, \quad \psi_i = \frac{x_i - x}{h}. \tag{3}$$

Common kernel functions include the Gaussian, $K(\psi_i) = \phi(\psi_i)$ where ϕ represents the standard normal density function; the Bisquare, $K(\psi_i) = (1 - \psi_i^2)^2 I(|\psi_i| < 1)$; and the Tricube, $K(\psi_i) = (1 - \psi_i^3)^3 I(|\psi_i| < 1)$, among others. In these expressions, I is a simple dummy variable that equals one when the condition is true. Kernel regression predictions are not sensitive to the choice of the kernel, but they are quite sensitive to the bandwidth or window width choice.[5]

In practice, nonparametric estimators can be estimated at any arbitrary set of target values; a common practice is to use each observation, in turn, as the target point. Upon hearing that separate kernel regressions may be estimated for each observation, people who have been trained in standard parametric modeling procedures often think that there must have been some sleight of hand that prevents all potential degrees of freedom from being exhausted in the

[4]Good, thorough reviews of nonparametric procedures are presented in Li and Racine (2007), Loader (1999), and Pagan and Ullah (1999). The discussion presented here draws heavily from Loader (1999), Pagan and Ullah (1999), Cleveland and Devlin (1988), and my own applications. Useful surveys include Härdle and Linton (1994) and Yatchew (1998).

[5]A "bandwidth" is a fixed value of h that does not vary depending on the point where the function is being evaluated. A "window size" is a bandwidth that varies across x. For example, we might choose to use the nearest 25 percent of the observations to construct the estimated value of y at x. When the values of x are distributed fairly evenly across its range of values, there is little difference between working with a fixed bandwidth and a comparable variable window size. However, the variable window approach is preferable when there are areas of x where the data are sparse because expanding the size of the window keeps a very small number of observations from receiving undue weight in the calculation of $\hat{y}(x)$.

calculation. However, nonparametric procedures impose sufficient continuity that it is often the case that only a few more degrees of freedom are used when compared with a simple linear regression. As the target changes from observation i to observation $i+1$, the weights may change very little, producing nearly identical values for $\hat{y}(x_i)$ and $\hat{y}(x_{i+1})$. The degrees of freedom used in estimating the n predicted values of y can be calculated by gathering all of the weights implicitly defined by Equation (3) into one large, $n \times n$ matrix, $\mathbf{L}$, and writing $\hat{\mathbf{Y}} = \mathbf{LY}$. The degrees of freedom used in estimation is then simply the trace of $\mathbf{L}$, i.e., $d = \text{tr}(\mathbf{L})$.

Kernel regressions can also be used to calculate the marginal effect of x at a given point. All that is necessary is to calculate the derivative of Equation (3) analytically at the target value of x. However, it turns out to be possible to estimate predicted values for the function and its derivative simultaneously using local linear regressions. Note that Equation (3) is equivalent to a weighted least squares regression of y on an identity vector with weights defined by $(K(\psi_i))^{1/2}$. The "locally weighted regression" (LWR) approach generalizes this idea by adding $x_i - x$ as an explanatory variable.[6] Thus, we are using a local linear approximation to predict the value of y at point x. Everything else is the same as before: to construct $\hat{y}(x)$, we first subtract the target value, x, from each value of x_i, and form the kernel, $K(\frac{x_i - x}{h})$. $\hat{y}(x)$ is the predicted value from the weighted least squares regression of y on a constant and $x_i - x$ with weights of $(K(\psi_i))^{1/2}$. The only difference from kernel regression is the addition of the explanatory variable to the constant term. $\hat{y}(x)$ is the predicted value of the regression at $x_i = x$, i.e., it is the intercept of the weighted least squares regression. The coefficient on $x_i - x$ is an estimate of the slope of the function at x. However, it is very important to recognize that the optimal bandwidth or window size is likely to be much larger when the objective is to estimate the marginal effect of x on y rather than to predict y directly. How much larger remains an open issue despite the voluminous literature on bandwidth selection.

LWR estimates of the FAR model are shown in Figure 2. I used the GCV criterion to pick the window size (Loader, 1999; McMillen and Redfearn, forthcoming). After varying h from 10 percent to 30 percent in increments of 1 percent, the lowest value of the GCV occurs at $h = 0.14$. At this value, $\text{tr}(\mathbf{L}) = 13.99$, i.e., approximately 14 degrees of freedom are used to predict the value of ln(FAR) at every data point. This figure compares with 5 degrees of freedom for the spline function and 13 for the Fourier expansion. Figure 2 suggests that the LWR estimator combines the best features of the other two approaches. It successfully tracks the sharp rise in ln(FAR) near the city center, while being less influenced by outliers at large distances. However, there is no particular reason to prefer any of the approaches to another. Since they are all easy to calculate, it is a good idea to compare the results for several choices.

[6]Locally weighted regression was developed by Cleveland and Devlin (1988) and was used first in the urban economics literature by Meese and Wallace (1991).

Spatial Lag Models

Although they are analyzed elsewhere in this special issue, some comment is in order here concerning spatial lag models, which are a commonly used alternative to the flexible functional form approaches considered so far. The two standard spatial models are the spatial autoregressive (AR) model, $\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\beta + \mathbf{u}$ and the spatial error model, $\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$, with $\mathbf{u} = \theta\mathbf{Wu} + \mathbf{e}$. Solving for $\mathbf{Y}$ the models can be written as follows:

$$\boldsymbol{Y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X}\beta + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{u}, \tag{4}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + (\boldsymbol{I} - \theta\boldsymbol{W})^{-1}\boldsymbol{e}. \tag{5}$$

$\mathbf{W}$ is an $n \times n$ weight matrix with (usually) prespecified weights. A common practice is to define $\mathbf{W}$ such that each tract bordering a zone on a map is given equal weight, with the weights summing to one. Thus, the first row of $\mathbf{W}$ might have four values of 0.25 with zeros elsewhere, and the second row might have five values of 0.20. Alternatively, $\mathbf{W}$ might be based on distance, e.g., giving each observation in a 1-mile radius equal weights that sum to unity, with all other values equaling zero.

Note the remarkable similarity between the specification of $\mathbf{W}$ and the kernel weight function. Indeed, the common approaches mentioned here are equivalent to a "rectangular" or "uniform" kernel with a very narrow bandwidth or window size. An important difference is that the spatial lag approach allows the researcher to test formally whether a single parameter—ρ or θ—equals zero, which is easier than formally testing for linearity as the null model in a nonparametric regression equation. However, the spatial lag model is built on a paradox. The whole reason for introducing the spatial weight matrix is that the base $\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$ model leaves unexplained spatial effects in the residuals, yet a parametric structure is being imposed on the data as though the true model structure were known beforehand. Both approaches require manipulation of $n \times n$ matrices, yet the models are typically estimated using maximum likelihood procedures that require large samples to be reliable.

Though the spatial AR model and the spatial error model often produce nearly identical results, the rationale behind them differs markedly. The spatial error model is typically viewed as a quick fix for autocorrelation, a "correction" when the residuals imply some remaining dependence even though the base equation appears reasonable. This approach is troublesome because it requires the researcher to specify the actual structure of the errors—the unknown part of the equation. The spatial error approach estimates a fully parametric model using maximum likelihood procedures, yet the reason that it is being used is that the true structure of the model is unknown. Again, flexible functional form and nonparametric methods can accomplish the same objective of reducing spatial autocorrelation without imposing arbitrary parametric structure.

Although the spatial AR model also is often treated as a quick fix for spatial autocorrelation, it actually makes more sense than the spatial error

model because it tries to specify how the dependent variable responds to its neighboring values.[7] Structure is clearly necessary to express the relationship between each of the n values of y and each of its $n - 1$ potential neighbors. If the goal is to test the significance of nearby values of the dependent variable on y, a parametric model such as Equation (4) is the only feasible way to proceed. But unless the objective is to estimate this causal relationship, there is no reason to impose arbitrary structure on the model simply because spatial autocorrelation is present in the residuals. Instead, further investigation is needed to determine the true structure of the model, or we should admit at the start that the true nature of the spatial model is unknown by using flexible functional forms or nonparametric procedures.

The last two columns of Table 1 present estimates of two spatial AR models. The first model has distance from the city center as its single explanatory variable, while the second uses the spline model as its base. The estimate of ρ is highly statistically significant in both cases, falling from a very high 0.833 to a still high value of 0.709 in the spline model.[8] Thus, spatial autocorrelation remains in the model even after adopting a highly flexible functional form.

How serious is this spatial autocorrelation problem? An advantage of this simple empirical application is that it allows us to analyze the data with simple graphs. Although there is some disagreement in the literature concerning the best way to base the predictions for a spatial lag model, for this application I used the deterministic portion of the right-hand side of Equation (4), $\hat{\boldsymbol{Y}} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{X}\beta$. These calculations are shown in Figure 3 for the linear and spline versions of the spatial lag model. Neither set of predictions appears much different than its counterpart regression with ρ set to zero (these regressions are not shown here to keep the figure simple). The primary difference between the spatial lag predictions and the comparable regression estimates is the small amount of noise around the clear trends. Figure 3 suggests that the spatial lag model's primary benefit is to capture some local variation in the dependent variable around its overall trend. Despite the apparently high degree of spatial autocorrelation, there appears to be little to be gained in analyzing floor area ratios using a spatial AR model when a simple series expansion or nonparametric approach clearly accounts well for the vast majority of the spatial variation in log(FAR).

---

[7]Bordignon, Cerniglia, and Revelli (2003), Brett and Pinkse (2000), Brueckner (1998), Brueckner and Saavedra (2001), Case, Rosen, and Hines (1993), Fredriksson and Millimet (2002), Millimet and Rangaprasad (2007), and Saavedra (2000) are examples of good applications of the spatial lag model. Brueckner (2006) provides a general framework for theory leading to a spatial lag model.

[8]For both models, I use first-order contiguity to define the spatial lag matrix for the census tracts. The models are estimated by maximizing the log-likelihood function implied by Equation (4) under the assumption that the underlying errors ($\mathbf{u}$) are independently and identically distributed normal.

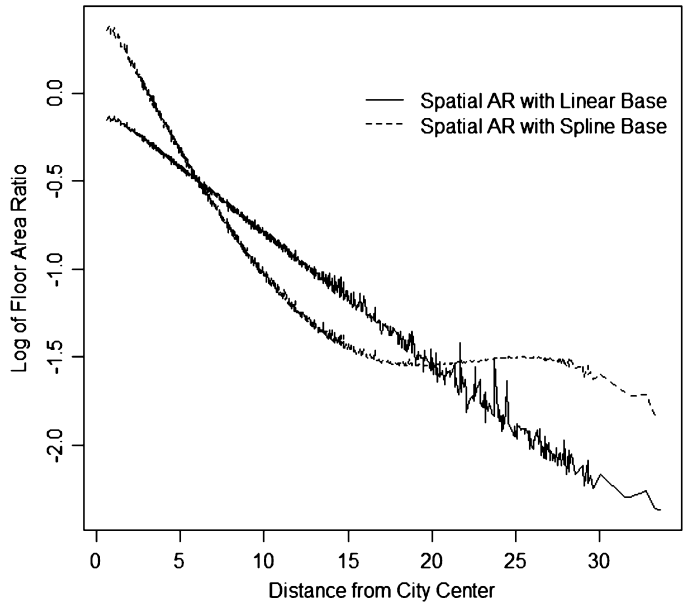

FIGURE 3: Spatial AR Predictions: $\hat{Y} = (I - \rho W)^{-1} X\beta$.

3. NONPARAMETRIC APPROACHES FOR MULTIVARIATE MODELS

Nonparametric approaches can easily be modified to take into account the local variation in ln(FAR) that is seen in the spatial AR estimates. All spatial variables are implicitly functions of the underlying geographic coordinates. Let *lo* and *la* denote the geographic coordinates, i.e., the longitude or latitude. The most common approach in a model with two explanatory variables is to use a simple product kernel such as $K_{lo}(\frac{lo_i - lo}{h_{lo}})K_{la}(\frac{la_i - la}{h_{la}})$, where *lo* and *la* are the target values and the subscript *i* indicates the observation, as before.[9] After constructing this bivariate kernel, the kernel regression estimates can be constructed as before, or the LWR version can be estimated using a weighted least squares regression of *y* on an intercept, $lo_i - lo$, and $la_i - la$. Researchers in urban economics, regional science, and geography have used a variant of LWR that has come to be known as "geographically weighted regression." In Figures 1 and 2, the base model takes the form $y_i = \beta_0 + \beta_1 x_i + u_i$; geographically weighted regression allows the coefficients β_0 and β_1 to vary spatially by writing them as

[9]The geographic coordinates could just as easily be the distances north and west of a given point, or polar coordinates could be used as in Cameron (2005). In practice, researchers typically use a single bandwidth or window size, *h*, in place of the two individual values. In order to do so, the variables must first be normalized by dividing by their standard deviations to remove scale differences.

functions of lo and la. The model becomes $y_i = \beta_0(lo_i, la_i) + \beta_1(lo_i, la_i)x_i + u_i$. In the statistical literature, this model is referred to as a "conditionally parametric model" because given the values of lo and la, the model is a simple parametric function of the primary explanatory variable.[10]

One problem that has not been addressed adequately in the literature is that x is itself a function of the geographic coordinates when it is a measure of proximity. Thus, the model can be written as $y_i = \beta_0(lo_i, la_i) + \beta_1(lo_i, la_i)x_i(lo_i, la_i) + u_i$ or simply $y_i = f(lo_i, la_i) + u_i$. As the bandwidth or window size decreases, there may be virtually no independent variation of x and the two geographic coordinates. Thus, small window sizes lead to imprecise estimates of the marginal effect of x. It may be that this issue is similar to the earlier argument that larger values of h are needed when the objective is to measure marginal effects rather than to simply construct predicted values for the dependent variable. More work is required on bandwidth and window size selection for spatial models in which the objective is to measure marginal effects of an underlying measure of proximity.

Semiparametric Regression

Nonparametric estimation suffers from a "curse of dimensionality"—the variance of the estimates increases rapidly with the number of variables. In this situation, semiparametric models become an attractive alternative to full nonparametric estimation. Semiparametric models separate the equation into parametric and nonparametric components, i.e.,

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{f}(\boldsymbol{Z}) + \boldsymbol{u}, \tag{6}$$

where $\mathbf{X}$ is a set of variables whose effects on $\mathbf{Y}$ are assumed to be modeled adequately using a simple parametric function and $\mathbf{Z}$ is a set of variables whose effects enter the equation nonparametrically. $\mathbf{Z}$ might be restricted to a single variable; it might include the geographic coordinates only; or it might simply include a subset of the full explanatory variable list. The advantage of the semiparametric approach is that it imposes parametric structure where the structure may be reasonable, while leaving the structure of the equation unrestricted for another set of variables. Hypothesis testing is easy because the parametric portion of the regression produces coefficient estimates and standard errors, and degrees of freedom are preserved by confining nonparametric

[10]The conditionally parametric model is considered in detail by Cleveland, Grosse, and Shyu (1992) and Cleveland (1994), while still more general versions are analyzed in Hastie and Tibshirani (1993). The spatial version of the model was first used by McMillen (1996), although the term "geographically weighted regression" was coined by Brunsdon, Charlton, and Fotheringham (1996). In practice, the kernel function for the spatial version of the conditionally parametric model is more often based on the simple distance between two points rather than the underlying geographic coordinates. McMillen and Redfearn (forthcoming) discuss the relationship among the various approaches in some detail.

modeling to **Z** alone. The semiparametric estimator involves no additional programming effort beyond that required to construct the simple nonparametric models discussed in previous sections.[11]

It is also possible to use flexible series expansions in place of the semiparametric model. Generalizations of the Fourier and spline approaches are available for two variables. Indeed, these approaches are probably preferable to a semiparametric approach if the goal is simply to allow for nonlinear effects in a single variable such as distance from the CBD. However, the semiparametric approach is a particularly easy and flexible approach for modeling broad spatial trends while also permitting the effects of other explanatory variables to vary by location.

4. FIXED EFFECTS VERSUS SEMIPARAMETRIC MODELS

The semiparametric approach can be used to control for general spatial trends in a model taking the form $\mathbf{Y} = \mathbf{X}\beta + f(\boldsymbol{la}, \boldsymbol{lo}) + \mathbf{u}$. Many researchers use fixed effects for municipalities or census tracts as controls for locations. The motivation for the fixed effects model is that there may be omitted, time-invariant variables that are correlated with the error term; if the variables are constant across all observations within a small geographic area, the fixed effects will control for the effects of the missing variables. The primary difference between the semiparametric and fixed effects approaches is whether omitted spatial effects can be expected to vary smoothly over space or have common values within discrete zones. The fixed effects approach will work well if the omitted variables comprise variables such as tax rates that do not vary within a district; it will not work as well if the omitted variables are measures of proximity that vary within districts.

Since a function can be approximated by a series of averages over very small intervals, it is often claimed that fixed effects for small zones may work well even if the actual spatial effects varies over space. The problem with this argument is that fixed effects are a profligate consumer of degrees of freedom. As an example, Table 2 presents selected results from a representative hedonic price function. The dataset includes all sales of small residential homes (six units or fewer) in Chicago for 1990–1991 and 1993–1999 (data for 1992 are unavailable), along with standard property characteristics.[12] I have included two variables of potential policy interest in the regression. The first, distance

[11]The semiparametric estimator is developed in detail in Robinson (1988). See McMillen and Redfearn (forthcoming) for more discussion of the advantages of the semiparametric approach.

[12]In addition to the year of sale, control variables include distance from the CBD, distance from Lake Michigan, building area, lot size, age, number of bedrooms; indicators that the property is near an el line or a rail line; and indicators that it has two or more stories, brick construction, a basement, an attic, central air conditioning, a one-car garage, a larger garage, or a fireplace. There are 82,807 sales in the sample. Thus, the spatial AR and spatial error models are not even feasible to estimate.

TABLE 2: House Price Regression Results for Selected Spatial Variables

	Parametric Regression: Fixed Effects			SemiParametric Regression: Window Size		
Explanatory Variable	None	Community Areas	Census Tracts	90%	50%	10%
Distance from El stop	0.070 (45.409)	0.045 (16.202)	0.013 (2.465)	0.051 (33.450)	0.040 (22.516)	0.021 (6.244)
Number of murders in census tract during year	−0.012 (172.885)	−0.001 (10.769)	−0.001 (6.624)	−0.007 (85.726)	−0.005 (63.579)	−0.002 (19.128)
Degrees of freedom used	27	102	790	27.30	31.73	62.47
R^2	0.552	0.737	0.788			

Notes: Absolute z-values are in parentheses below the estimated coefficients. The number of observations is 82,807. The regressions also include controls for nine years of sale, distance from the CBD, distance from Lake Michigan, building area, lot size, age, number of bedrooms; indicators that the property is near an EL line or a rail line; and indicators that it has two or more stories, brick construction, a basement, an attic, central air conditioning, a one-car garage, a larger garage, or a fireplace.

to the nearest stop on an elevated train line (the "EL"), is explicitly spatial.[13] In addition, I have included a variable—the number of murders in a census tract in a year—meant to be representative of the type of question often addressed using hedonic analysis: how much will households pay for a reduction in the number of homicides in the area? Although this variable is not explicitly spatial, crime rates are clearly higher in some areas than in others.[14]

Table 2 compares the results of parametric regressions with spatial fixed effects and semiparametric regressions with the geographic coordinates in the nonparametric part of the model. I use two alternative definitions of the fixed effects. The "community area" is a neighborhood definition for Chicago dating to studies done by University of Chicago sociologists in the 1930s. Although the community areas are large and comprise some heterogeneous areas, they are still in common use both by academics and by real estate agents, and each community area is sufficiently large to include a good number of observations. In contrast, census tracts are so small that many had no sales of small residential properties during this time. For the semiparametric models, I use a tri-cube product kernel and a LWR specification for the nonparametric portions of the estimation procedure.

The fixed effects estimator becomes a very blunt instrument when the geographic unit is small. Since identification comes from the deviations of

[13]Many authors have used hedonic studies to measure the benefits of proximity to transit lines; good examples include Baum-Snow and Kahn (2000), Bowes and Ihlanfeldt (2001), Gibbons (2004), and McMillen and McDonald (2004).

[14]Examples of studies analyzing the effect of crime on property values include Gibbons and Machin (2005), Pope (2008), Pope and Pope (2009), and Schwartz, Susin, and Voicu (2003).

the individual values from their group means and there may be virtually no variation in spatial variables within groups, it should not be surprising to find that the coefficients are quite sensitive to the number of fixed effects. The *z*-values drop dramatically for a variable like distance from the nearest EL stop when the number of fixed effects increases from 102 to 790—precisely what would be expected from a variable that has very little variation within an area as small as a census tract. The murder variable is problematic because it is measured at the census tract level. As the unit for the fixed effect decreases in size to the level of the census tract, identification is driven entirely by the variation within a census tract over time. Many census tracts have no murders; these offer no information to the tract-level fixed effects estimator. The result is that standard errors rise dramatically when the number of fixed effects increases.

The semiparametric estimator has a similar pattern in that the standard errors increase as the window size declines. Table 2 shows the implicit number of degrees of freedom used in estimation by the semiparametric regressions. Interestingly, even the quite small 10 percent window uses far fewer degrees of freedom to approximate the spatial effects than is the case with either fixed effects specification. A significant advantage of the semiparametric approach is that it greatly facilitates robustness checks: by varying the window size, it is easy to assess the sensitivity of the results to local geographic effects.

Perhaps surprisingly, the coefficient for distance from the nearest EL stop indicates that house prices are higher *farther* from the EL—the opposite of the result found in McMillen and Redfearn (forthcoming). The reason is simple: Table 2 uses data from the entire city of Chicago, whereas the other studies use only sales from areas close to the EL lines. The effect of distance from the EL is actually confined to relatively small areas near the lines; more distant locations can have quite high property values that have nothing to do with access to rapid transit. Thus, the apparent positive effect decreases substantially as the window size declines or when census tracts for the fixed effects estimator. The value of proximity can vary dramatically over space. When analyzing an extremely large city, it may be preferable to estimate separate models for different areas rather than try to model the entire city at once.[15]

[15]Using nonparametric approaches, McMillen and Redfearn find that proximity to an EL stop leads to significant increases in property values on the North and Southwest sides of the city. They argue that "In lower-income neighborhoods on the west and south sides of the city, proximity to the EL appears to be a disamenity. In these neighborhoods, the EL is frequently lined by vacant lots and failing businesses, which may lead to the positive correlation between house prices and distance from the EL. Thus, the nonparametric LWR estimates reveal a source of model misspecification that would be far from obvious in a more traditional parametric estimation strategy" (McMillen and Redfearn, forthcoming, p. 3). Redfearn (2009) also finds substantial heterogeneity in the effects of access to rapid transit stops on house prices in Los Angeles, with apparently positive effects in some areas and negative effects in others. As he points out, a combination of positive and negative results suggests that "amenity value is in the eye of the modeler" (Redfearn, 2009, p. 302).

This example illustrates clearly the fundamental problem with spatial data analysis for multivariate models: no matter how many variables are included in a model, ultimately the model is a function of only two—the geographic coordinates. This problem should be self-evident for variables like distance from an EL stop. It is less obvious for variables like crime rates, although they clearly have a strong spatial component. The issue even arises for the housing characteristics variables because certain areas of the city are more likely to have small lots, old houses, brick homes, fireplaces, and so on. In the end, the independent effects of various spatial variables are identified only through functional form assumptions.

To a certain extent, this observation is trivial—after all, distance from the CBD is one function of the geographic coordinates, while distance from an EL stop is another function of the same variables. But this seemingly trivial point has significant implications. First, the results of models relying heavily on pre-imposed parametric structure should be viewed with a great deal of skepticism. Thus, fixed effects and semiparametric models are always preferable to a spatial error model, and are also preferable to a spatial AR model unless the objective is explicitly to model the effects of spatial lags of the dependent variable on itself. Second, it is important to assess the sensitivity of the results to changes in the assumed model structure. Varying window sizes or the number of fixed effects, allowing for nonlinearities, and perhaps omitting some of the explicitly spatial variables can help assess the robustness of the results.

5. DISCRETE ZONES AND TREATMENT EFFECTS

Spatial data analysis can be simpler when geographic zones are discrete. For example, consider Holmes' (1998) influential paper analyzing the effect of "right-to-work" laws on the manufacturing share of total employment in U.S. counties. A right-to-work state is generally considered to have a more pro-business environment because it has adopted legislation banning requirements that all workers join a union in order to work at a firm. Holmes compares the employment shares for counties on either side of state borders for right-to-work states and those without right-to-work legislation. Since bordering counties are presumably the same in other ways—similar proximity to the transportation network, similar demographics, the same climate, etc.—the idea is that most of the determinants of employment shares will be the same on either side of state borders. To quote Holmes (1998, p. 668), "if state policies are an important determinant of the location of manufacturing, one should find an abrupt change in manufacturing activity when one crosses a border at which policy changes, because state characteristics unrelated to policy are the same on both sides of the border." Similar estimation strategies have been used by such authors as Black (1999) to measure the effect of school quality on house prices, and Cunningham (2007) and Zhou, McMillen, and McDonald (2008) to analyze the effect of zoning policies on land values.

We have already seen that discrete geographic zones can simplify the analysis of spatial data since fixed effects can control for missing variables that are constant within clearly defined zones. Previously, I suggested that a limitation of the fixed effects approach is that broad spatial trends vary within zones as well as across them. However, school districts and zoning boundaries are examples of well-defined districts that might be expected to share common fixed effects for observations within each district. The problem of missing, spatially correlated data is not necessarily eliminated by the introduction of even an appropriately defined fixed effect, however. For example, consider Black's (1999) analysis of the effect of school quality on house prices. Since school quality is presumably the same for every student attending the school, fixed effects can control adequately for school quality. By comparing house prices on either side of a school district boundary, she hopes to "effectively remove the variation in neighborhoods, taxes, and school spending" (Black, 1999, p. 577). Nevertheless, it still is necessary to control for differences in housing characteristics and within-district variation in such variables as access to transportation or the presence of parks. Since these variables are likely to share similar values within districts, the school district fixed effect may be correlated with the error term in the hedonic price function. This correlation leads to biased estimates of the effect of school quality—i.e., the district fixed effect—on house prices if data are pooled across districts.

Black's insight is that homes on either side of a district boundary are likely to have been built at the same time, have similar structural characteristics, and have virtually identical access to transportation or parks. Thus, limiting the sample to sales taking place close to the district boundaries can control for the effect of any missing variables. In effect, homes along one side of the boundary share a common fixed effect even if homes within the same district differ in unmeasured ways when they are farther from the boundary. To draw an analogy to nonparametric estimation, the boundary approach compares the predicted values from a narrow window of observations on either side of the boundary.

Although this approach helps to reduce the effects of missing spatial variables, it may not eliminate the problem altogether. For example, suppose that homes on one side of the district boundary were poorly built or since have been poorly maintained. Unless adequate controls are included for these effects, the analysis may spuriously indicate that low-quality schools lead to low prices when the actual problem is that low-quality housing leads to low prices. Even if the indication of high prices is not spurious, it is likely to be overstated if the difference between homes is not captured by the explanatory variables included in the regression. Although the approach may mitigate the effect of missing variables, it still requires that spatial effects be confined to simple discrete changes at district boundaries. Spatial variation within zones can severely bias the results.

The endogeneity of boundary districts is another problem that potentially arises when analyzing data from discrete zones. In the case of land-use zoning,

for example, it is distinctly possible that an area that is particularly well suited for an office building will be zoned for commercial rather than residential use. A common procedure for estimating the causal effect of zoning on sales prices is to regress prices on a host of variables—property characteristics such as square footage; controls for location such as distance to highway interchanges—while including zoning as an explanatory variable. But if properties are likely to be zoned for the their "highest and best use"—i.e., the use that leads to the highest land value—then missing variables that are correlated with zoning status are also correlated with sales prices, which leads to biased estimates of the effect of zoning on sales prices.

Zoning is another example of situation where a boundary sampling style can reduce the bias: the difference in sales prices across properties on either side of a zoning boundary will provide a clean measure of the causal effect of zoning on prices if the properties are identical in all other ways. The properties are nearly certain to be identical in terms of location if they are paired directly across boundaries; the issue is whether adequate controls are available for structural characteristics. In the case of undeveloped, vacant land (e.g., Cunningham, 2007; Zhou et al., 2008) a comparison of prices of boundaries should control very well for omitted variable that determine zoning status as well as omitted variables that influence sales prices in other ways. The approach will not work as well for sales prices of properties that include structures. Also, if we compare zoning status across hundreds or thousands of property pairs at zoning boundaries throughout a large city, there may be some areas where zoning is binding and others where zoning has no effect on sales prices. Thus, there may still be reason to include controls for broader spatial effects.

Comparing sales prices for narrow windows across district boundaries is a special case of matching estimators.[16] The general idea behind a matching estimator is to balance the distribution of the covariates so that they are the same across two samples. In the case of land-use zoning, our objective is to infer an unobserved event—the price of a residential property if it were zoned for commercial use or the price of a commercial property if it were zoned for residential use. If we had the ability to randomly assign properties to one category, a simple difference in the mean values of the properties would be sufficient to measure the effect of zoning on property values. In this case of an idealized natural experiment, we would not expect to observe any differences in the overall distributions of standard explanatory variables across zoning categories. Of course, individual properties would be in different locations and have different structural characteristics, but these differences would be "averaged out" over a large sample of randomly assigned properties.

[16]The literature on matching estimators has grown quite large. Good starting points include the February 2004 special issue of the *Review of Economics and Statistics* on "The Econometrics of Matching," the March–April 2005 special issue of the *Journal of Econometrics* on "Experimental and Non-Experimental Evaluation of Economic Policy and Models," and Volume 6B of the *Handbook of Labor Econometrics* (Heckman and Leamer, 2007).

The key insight behind a matching estimator is that the requirements for estimating average treatment effects are different from the requirements for obtaining unbiased estimates of all of the coefficients in a structural equation. If our objective is to estimate the effect of zoning on property values, we may not care much about the marginal return of an additional bedroom to the sales price of property. In this case, we can compare mean values across matched samples to estimate a direct causal effect. The boundary estimator accomplishes the matching by comparing properties on either side of the border. This matching procedure controls for location, but does not necessarily control adequately for housing characteristics; hence housing characteristics continue to be included as explanatory variables. More complicated multivariate metrics can match properties by structural characteristics in addition to location. Successful matching produces balanced sample of properties that mimic the properties of a natural experiment; any difference between the distributions of explanatory variables across districts is essentially random.

The propensity score approach is a special case of the matching estimator that reduces the metric for identifying matches to a simple comparison of the predicted values from a probit or logit model that explains the probability that an observation has received treatment. Note the direct link between the probit/logit model and the issue of endogeneity. In the case of zoning, we begin by estimating a model explaining the probability that a parcel has been zoned for commercial use. We then match each commercial parcel with a residential parcel that has a nearly identical probability of being zoned for commercial use. Alternatively, we could direct match parcel based on their observed location and structural characteristics. In the end, a simple comparison of mean values can potentially be sufficient to estimate the average effect of zoning on property values.[17]

6. CONCLUSION

My focus in this paper has been on cross-sectional spatial data analysis. I argue that the main difficulty with modeling spatial data is a combination of nonlinearity and missing variables that are correlated over space. While spatial lag models attempt to account for these problems using conventional parametric model specifications, they become unwieldy in large datasets and rest on an unreasonable assumption that the true model structure is known beforehand. Since the reason for considering a spatial lag model in the first place is that spatial autocorrelation remains in the model after an apparently reasonable parametric structure has been assumed, it makes more sense to admit at the onset that the true model structure is not known.

[17]The propensity score approach was originally proposed by Rosenbaum and Rubin (1983, 1984). Early examples of their use in urban economics include McMillen and McDonald (2002) and Reed and Rogers (2003).

Given this somewhat pessimistic view of our ability to accurately model spatial data, it becomes critical to subject estimated models to a serious of specification tests in order to assess the robustness of the results. Spatial lag models are a useful tool for testing model specifications. Series expansions and nonparametric estimators are flexible approaches that have enormous advantage when analyzing spatial data. While often using fewer degrees of freedom than standard fixed effects estimators, they also can be adapted to allow for both continuous and discrete spatial trends in both the dependent variable and the marginal effects of the explanatory variables. The key is to avoid becoming tied to a single modeling approach, to become familiar with techniques used in other fields, and to think of econometric modeling as a means of testing the robustness of a model specification.

This focus on model specification is less important when the objective is to measure average treatment effects. It may not be necessary to correctly specify a full hedonic housing if the objective is to measure the average effect of a treatment that is confined to a well-defined geographic areas. Comparisons of average prices on either side of the border may provide good estimates of the treatment effect. However, these approaches still face potential problems with omitted variables since spatial effects do not necessarily match district boundaries perfectly.

REFERENCES

Ahlfeldt, G. M. and N. Wendland. 2008. *Fifty Years of Urban Accessibility: The Impact of Urban Railway Network on the Land Gradient in Industrializing Berlin*. Zurich: Swiss Economic Institute.

Alonso, W. 1964. *Location and Land Use*. Cambridge, MA: Harvard University Press.

Anderson, J. E. 1982. "Cubic-Spline Urban-Density Functions," *Journal of Urban Economics*, 12, 155–267.

———. 1985. "The Changing Structure of a City: Temporal Changes in Cubic Spline Urban Density Functions," *Journal of Regional Science*, 25, 413–425.

Atack, J. and R. A. Margo. 1998. "Location, Location, Location! The Price Gradient for Vacant Urban Land: New York, 1835 to 1900," *Journal of Real Estate Finance and Economics*, 16, 151–172.

Baum-Snow, N. and M. E. Kahn. 2000. "The Effects of New Public Projects to Expand Urban Rail Transit," *Journal of Public Economics*, 77, 241–362.

Black, S. E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114, 577–599.

Bordignon, M., F. Cerniglia, and F. Revelli 2003. "In Search of Yardstick Competition: A Spatial Analysis of Italian Municipality Property Tax Setting," *Journal of Urban Economics*, 54, 199–217.

Bowes, D. R. and K. R. Ihlanfeldt. 2001. "Identifying the Impacts of Rail Transit Stations on Residential Property Values," *Journal of Urban Economics*, 50, 1–25.

Brett, C. and J. Pinkse. 2000. "The Determinants of Municipal Tax Rates in British Columbia," *Canadian Journal of Economics*, 33, 695–714.

Brueckner, J. K. 1998. "Testing for Strategic Interaction Among Local Governments: The Case of Growth Controls," *Journal of Urban Economics*, 44, 438–467.

———. 2006. "Strategic Interaction Among Governments," in R. J. Arnott, and D. P. McMillen (eds.), *A Companion to Urban Economics*. Malden, MA: Blackwell, pp. 332–347.

Brueckner, J. K. and L. A. Saavedra. 2001. "Do Local Governments Engage in Strategic Property-Tax Competition?" *National Tax Journal*, 54, 203–229.

Brunsdon, C., A. S. Fotheringham, and M. Charlton. 1996. "Geographically Weighted Regression," *Geographical Analysis*, 28, 281–298.
Cameron, T. A. 2005. "Directional Heterogeneity in Distance Profiles in Hedonic Property Value Models," *Journal of Environmental Economics and Management*, 51, 26–45.
Case, A. C., H. S. Rosen, and J. R. Hines. 1993. "Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States," *Journal of Public Economics*, 52, 285–307.
Clark, C. 1951. "Urban Population Densities," *Journal of the Royal Statistical Association Series A*, 114, 490–496.
Cleveland, W. S. 1994. "Coplots, Nonparametric Regression, and Conditionally Parametric Fits," in T. W. Anderson, K. T. Fang, and I. Olkin (eds.), *Multivariate Analysis and its Applications*. Hayward, CA: Institute of Mathematical Statistics, pp. 21–36.
Cleveland, W. S. and S. J. Devlin. 1988. "Locally Weighted Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596–610.
Cleveland, W. S., E. H. Grosse, and W. M. Shyu. 1992. "Local Regression Models," in J. M. Chambers, and T. J. Hastie (eds.), *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks/Cole, pp. 309–376.
Coulson, N. E. 1991. "Really Useful Tests of the Monocentric City Model," *Land Economics*, 67, 299–307.
Cunningham, C. R. 2007. "Growth Controls, Real Options, and Land Development," *Review of Economics and Statistics*, 89, 343–358.
Fredriksson, P. G. and D. L. Millimet. 2002. "Strategic Interaction and the Determinants of Environmental Policy Across U.S. States," *Journal of Urban Economics*, 51, 101–122.
Gallant, R. 1981. "On the Bias in Flexible functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211–245.
———. 1982. "Unbiased Determination of Production Technologies," *Journal of Econometrics*, 20, 285–323.
Gibbons, S. 2004. "The Costs of Urban Property Crime," *Economic Journal*, 114, 441–463.
Gibbons, S. and S. Machin. 2005. "Valuing Rail Access Using Transport Innovations," *Journal of Urban Economics*, 57, 148–169.
Härdle, W. and O. B. Linton 1994. "Applied Nonparametric Methods," in R. F. Engle, and D. L. McFadden (eds.), *Handbook of Econometrics*, Vol. 4. New York: North-Holland, pp. 2295–2339.
Hastie, T. and R. Tibshirani 1993. "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Series B*, 55, 757–796.
Heckman, J. J. and E. E. Leamer. 2007. *Handbook of Econometrics*, Vol. 6b. New York: North-Holland.
Holmes, T. J. 1998. "The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders," *Journal of Political Economy*, 106, 667–705.
Ihlanfeldt, K. R. 2004. "The Use of an Econometric Model for Estimating Aggregate Levels of Property Tax Assessment Within Local Jurisdictions," *National Tax Journal*, 57, 7–24.
Li, Q. and J. S. Racine. 2007. *Nonparametric Econometrics*. Princeton, NJ: Princeton University Press.
Loader, C. 1999. *Local Regression and Likelihood*. New York: Springer.
McMillen, D. P. 1996. "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach," *Journal of Urban Economics*, 40, 100–124.
———. 2006. "Testing for Monocentricity," in R. J. Arnott and D. P. McMillen (eds.), *A Companion to Urban Economics*. Malden, MA: Blackwell, pp. 128–140.
McMillen, D. P. and J. Dombrow. 2001. "A Flexible Fourier Approach to Repeat Sales Price Indexes," *Real Estate Economics*, 29, 207–225.
McMillen, D. P. and J. F. McDonald. 2002. "Land Values in a Newly Zoned City," *Review of Economics and Statistics*, 84, 62–72.
———. 2004. "Reaction of House Prices to a New Rapid Transit Line: Chicago's Midway Line," *Real Estate Economics*, 32, 463–486.
McMillen, D. P. and C. Redfearn. 2010. "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions," *Journal of Regional Science*, forthcoming.

Meese, R. and N. Wallace. 1991. "Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices," *Journal of the American Real Estate and Urban Economics Association*, 19, 308–332.

Millimet, D. L. and V. Rangaprasad. 2007. "Strategic Competition Amongst Public Schools," *Regional Science and Urban Economics*, 37, 199–210.

Mills, E. S. 1969. "The Value of Urban Land," in H. Perloff (ed.), *The Quality of the Urban Environment*. Baltimore, MD: Resources for the Future, Inc., pp. 231–253.

———. 1972. *Studies in the Structure of the Urban Economy*. Baltimore, MD: Johns Hopkins Press.

Muth, R. F. 1969. *Cities and Housing*. Chicago, IL: University of Chicago Press.

Pagan, A. and A. Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Pope, J. C. 2008. "Fear of Crime and Housing Prices: Household Reactions to Sex Offender Registries," *Journal of Urban Economics*, 64, 601–614.

Pope, D. G. and J. C. Pope 2009. *Crime and Property Values: Evidence from the 1990's Crime Drop*. Blacksburg, VA: Virginia Tech University.

Redfearn, C. L. 2009. "How Informative Are Average Effects? Hedonic Regression and Amenity Capitalization in Complex Urban Housing Markets," *Regional Science and Urban Economics*, 39, 297–306.

Reed, W. R. and C. L. Rogers 2003. "A Study of Quasi-Experimental Control Group Methods for Estimating Policy Impacts," *Regional Science and Urban Economics*, 33, 2–25.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Rosenbaum, P. R. and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

———. 1984. "Reducing Bias in Observational Studies using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

Saavedra, L. A. 2000. "A Model of Welfare Competition with Evidence From AFDC," *Journal of Urban Economics*, 47, 248–279.

Schwartz, A. E., S. Susin, and I. Voicu. 2003. "Has Falling Crime Driven New York City's Real Estate Boom?" *Journal of Housing Research*, 14, 101–135.

Suits, D. B., A. Mason, and L. Chan. 1978. "Spline Functions Fitted by Standard Regression Models," *Review of Economics and Statistics*, 60, 132–139.

Thorsnes, P. and J. W. Reifel. 2007. "Tiebout Dynamics: Neighborhood Response to a Central-City/Suburban House-Price Differential," *Journal of Regional Science*, 47, 693–719.

Thorsnes, P., R. Alexander, and B. McLennan. 2009. *Low-Income Housing Built in High-Amenity Area: Long Run Housing-Market Effects of Exogenous Amenities*. Dunedin, New Zealand: University of Otago.

Yatchew, A. 1998. "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36, 669–721.

Zhou, J., D. P. McMillen, and J. F. McDonald. 2008. "Land Values and the 1957 Comprehensive Amendment to the Chicago Zoning Ordinance," *Urban Studies*, 45, 1647–1661.