Spatial Data Analysis: Specification Testing with Unknown Functional Form and Spatially Correlated Missing Variables

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
mcmillen@illinois.edu

April 15, 2009

Abstract

Nonlinearity and missing variables that are correlated over space pose serious problems for spatial data analysis. Imposing less structure than standard spatial lag models while being more amenable to large data sets, non-parametric and semi-parametric methods offer significant advantages for spatial modeling. While often using fewer degrees of freedom than standard fixed effects estimators, non-parametric methods 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. Since it is unlikely that the researcher can specify the true structure of a model, it is critical that spatial effects be tested for robustness.

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, the 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or is it the fact that the area has poor access to areas with growing employment and is filled with badly maintained, outdated housing?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 extraordinary difficult task.

In addition to functional form issues, the other central problem 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 data set 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. No matter how long the variable list gets, some variable is always missing. 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.

This 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 non-parametric 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 is omitted from the regression, no amount of statistical investigation will uncover 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 model 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 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. Non-parametric and semi-parametric models form attractive alternatives to parametric alternatives because they admit at the start that the true model structure is unknown. Unfortunately, the voluminous statistical literature on non-parametric and semi-parametric 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 non-parametric regression procedures. By focusing on this special case, the advantages of other estimators have been neglected. In addition, many researchers fail to understand that non-parametric 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 predicted values of the dependent variable and the marginal effects of the explanatory variables.

My objective in this paper is to provide an introduction to flexible functional forms and non-parametric alternatives to conventional parametric models suited to spatial data analysis. I begin by showing how flexible parametric functional forms and non-parametric alternatives can be used fruitfully in a simple single-explanatory variable setting. Next, I show how semi-parametric models can be used to combine standard hypothesis and robustness checks for multiple explanatory variable models with very large samples. In both of these settings, non-parametric models are likely to be superior to standard spatial econometric models. In addition, I argue that the standard approach of using fixed effects to control for location in a multiple explanatory variable model is likely to use an excessive number of degrees of freedom while providing less accurate results than non-parametric procedures. I am not arguing that non-parametric procedures are the only right way to go about conducting spatial data analysis. Rather, my objective is to show that they are a useful approach for conducting hypothesis testing and robustness checks when the true model is not known.

2. Single-Explanatory Variable Model: Series Expansions

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 include Abelson (1997), Ahlfeldt and Wendland (2008), Anderson (1982, 1985), Atack and Margo (1998), Coulson (1991), McDonald (1989), 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 data set here to illustrate some basic tools for spatial data analysis. As in the previous paper, the basic data set includes all 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

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

explanatory variable for 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.[2]

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 $y = \ln(FAR)$ on x. As expected, this simple linear model fits the data well, with distance alone accounting for 61.8% of the spatial variation in ln(FAR). The estimated coefficient implies that floor area ratios decline by 7.2% with each mile from the city center.[3] The predicted values from the regression form the straight line shown in Figure 1. The true relationship is clearly nonlinear, with a steeper gradient near the CBD and a flatter gradient at more distant locations.

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 et al, 1978) and the flexible fourier form (Gallant, 1981, 1992). 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

[2] GIS programs make these calculations easy by providing geographic coordinates for the centroids of map layers such as census tracts. By my calculations, Chicago's CBD has an approximate longitude (in radians) of lo1 = -2π87.627800/360 and a latitude of la1 = 2π41.881998/360. Let lo2 and la2 represent the geographic coordinates of a census tract centroid, again expressed in radians. Using a standard geographic formula, the distance between the census tract and the CBD point is acos(sin(la1)*sin(la2) + cos(la1)*cos(la2)*cos(lo2-lo1))*3958 miles.

[3] 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^2s 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.

intermediate intervals: $x_1 = x_0 + (x_S-x_0)/S$, $x_2 = x_0 + 2(x_S-x_0)/S$, …, $x_{S-1} = x_0 + (S-1)(x_S-x_0)/S$. Associated with each knot is a dummy variable D_s indicating whether x is greater than x_s. The estimating equation is

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

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:

$$y = \beta_0 + \beta_1 z + \beta_2 z^2 + \sum_{j=1}^{J}(\sin(jz) + \cos(jz)) + \mathrm{u} \qquad (2)$$

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).

In a single-explanatory variable model, simple visual inspection of the raw data and predictions can be sufficient to determine the order of the expansions, S for the spline function and J for the fourier expansion. More rigorous approaches can also be used. The most popular are the Akaike and Schwarz information criteria, both of which impose a penalty on adding explanatory variables to a model. The estimated variance for either equation (1) or (2) is $s^2 = n^{-1}\sum e_i^2$, where n is the number of observations and e_i is the residual for observation i. The spline function has 4+s estimated coefficients while the fourier expansion has 3+2J. Denoting the number of estimated coefficients by k, the two criteria are $AIC = \log(\hat{\sigma}^2) + 2k/n$ and $SC =$

$\log\left(\hat{\sigma}^2\right) + \ln(n)k/n$. Optimal values for S and J can be found by minimizing these criteria with respect to the number of expansion terms. Since 2/n < ln(n)/n, the AIC criterion will never indicate a higher value for k (and hence S or J) than the SC criterion. In practice, they often indicate the same values.

Spline and fourier expansion estimates of the FAR model are shown in Table 1. The number of expansion terms is S = 1 and J = 5 using the AIC criterion.[4] The R^2s indicate that at least 78% of the variance in ln(FAR) is explained by simple functions of distance from the CBD. The predicted values are shown in Figure 2. 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 data set, 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 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.

3. Single-Explanatory Variable Model: Kernel Regression

Non-parametric 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 pre-defined points.[5] Perhaps the most commonly used non-parametric approach across all fields is kernel regression. Letting x be any arbitrary value of the explanatory variable,

[4] The Schwarz criterion indicates optimal values of S = 0 and J = 1.

[5] Good, thorough reviews of nonparametric procedures are presented in Loader (1999), Li and Racine (2007), 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).

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, $K(\psi_i) = \left(1 - \psi_i^2\right)^2 I\left(|\psi_i| < 1\right)$; and the Tricube, $K(\psi_i) = \left(1 - \psi_i^3\right)^3 I\left(|\psi_i| < 1\right)$, 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. 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% of the observations to construct the estimated value of y at x. Thus, the window size for target point x, h(x), is simply the 25th percentile value of $|x_i - x|$. A simple rule of thumb that serves as a useful starting point for the bandwidth is to choose h = 1.06sn$^{-1/5}$, where s is the standard deviation of x and n is the number of observations, while the 25th percentile is a good starting point for the window size. 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 $\hat{y}(x)$.

Cross-validation is perhaps the most commonly used approach for choosing the bandwidth or window size. The usual approach is to calculate $\hat{y}(x)$ using every value of x in turn as the target value for the kernel regression. However, observation i is omitted from the kernel regression when calculating the predicted value at that point. The residual for observation i is then $e_i = y_i - \hat{y}(x_i)$, and the full residual sum of squares is $CV = \sum_{i=1}^{n} e_i^2$. The optimal value for the bandwidth or window size is the value that minimizes this residual sum of squares. Note that if observation i were not omitted from that observation's kernel regression the optimal value for h would shrink toward zero: any local regression with a single explanatory variable has a perfect fit within sample if the window is limited to two observations. Also, note that this calculation is computer intensive since separate kernel regressions are calculated for each data point and for each value of the bandwidth or window size.

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 calculation. However, non-parametric 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 regression 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, nxn matrix, L, and writing $\hat{Y} = LY$. Thus, the first row of L has the weights applied to each observation when calculating the predicted value of y at observation 1, the second row has the weights for observation 2, and so on. The degrees of freedom used in estimation is then simply the trace of L, i.e., d = tr(L). Loader (1999) uses this degrees of freedom calculation to construct a simple alternative to the cross validation criterion,

$$GCV = n\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}(x_i)\right)^2}{(n-d)^2} \tag{4}$$

where, unlike the usual CV measure, the predicted values $\hat{y}(x_i)$ are calculated without omitting observation i from the kernel regression.[6]

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" approach generalizes this idea by adding x_i-x as an explanatory variable.[7] 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, $K\left(\frac{x_i - x}{h}\right)$. $\hat{y}(x)$ is the predicted value from the weighted least squares regression of y on a constant and x_i-x with weights of

[6] The GCV criterion can also be used to pick the optimal expansion length for spline functions or fourier expansions. For these parametric models, d is simply the number of explanatory variables in the regression (including the intercept term).

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

$(K(\psi_i))^{1/2}$. 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 the 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 unclear at this point; twice the value indicated by the CV or GCV criterion is a reasonable starting point.

LWR estimates of the FAR model are shown in Figure 2. I used the GCV criterion to pick the window size. After varying h from 10% to 30% in increments of 1%, the lowest value of the GCV occurs at h = .14. At this value, tr(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 over another. Since they are all easy to calculate, it is a good idea to compare the results for several choices.

4. 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

AR model, Y = ρWY + Xβ + u and the spatial error model, Y = XB + u, with u = θWu + e. Solving for Y the models can be written as follows:

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} u \qquad (5)$$

$$Y = X\beta + (I - \theta W)^{-1} e \qquad (6)$$

W is an nxn weight matrix with (usually) pre-specified weights. A common practice is to define 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 W might have four values of 0.25 with zeros elsewhere, and the second row might have five values of 0.20. Alternatively, W might be based on distance, e.g., giving each observation in a one-mile radius equal weights that sum to unity, with all other values equaling zero.

Note the remarkable similarity between the specification of 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 non-parametric 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 Y = XB + 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 nxn 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 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. 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.[8] 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 (6) 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 non-parametric procedures.

[8] 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.

Table 2 presents 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. [9] Thus, spatial autocorrelation remains in the model even after adopting a highly flexible functional form.

How serious is this spatial autocorrelation problem? An advantage this simple empirical application is that it allows us to analyze the data with simple graphs. There is some disagreement in the literature concerning the best way to base the predictions in the spatial lag model. One possibility is simple to look at the "trend" – the estimated values of XB, ignoring the contiguity matrix. In his excellent "SPDEP" package written for the software program "R", Roger Bivand suggests basing the predictions on a combination of the trend and the "signal" component. The signal could be calculated as $\hat{\rho}WY$, but since this involves making predictions of the dependent variable based on its own values, Bivand suggests using a spatially lagged version in its place using the relationship $\hat{\rho}WY = \hat{\rho}W\left(I - \hat{\rho}W\right)^{-1} X\hat{\beta}$. The estimated trend and the combined trend and "feasible signal" are shown in Figure 3 and 4 for the linear and spline versions of the spatial lag model. Either one predicts Y well, but neither trend is an accurate characterization of an overall trend that clearly is close to the LWR estimates, which also are shown in Figure 3 and 4. Though the overall prediction is good, the decomposition of the prediction into trend and signal components does not appear useful in this example.

Another possible basis for predictions in the spatial lag model is the deterministic portion of the right hand side of equation (5), $\hat{Y} = \left(I - \rho W\right)^{-1} X\beta$. These calculations are shown in

[9] 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 (5) under the assumption that the underlying errors (u) are independently and identically distributed normal.

Figure 5 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. Together, Figures 3-5 suggest that the spatial lag model's primary benefit is to capture some local variation in the dependent variable around its overall trend.

5. Multivariate Locally Weighted Regression

Non-parametric approaches can easily be modified to take into account the local variation in ln(FAR), i.e., the "signal" component evident in Figure 3 and 4. 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.[10] The most common approach in a model with two explanatory variables is to use a simple product kernel: $K\left(\frac{lo_i - lo}{h_{lo}}, \ \frac{la_i - la}{h_{la}}\right) = K_{lo}\left(\frac{lo_i - lo}{h_{lo}}\right)K_{la}\left(\frac{la_i - la}{h_{la}}\right)$, where lo and la are the target values and the subscript i indicates the observation, as before. Any kernel listed above can be used; the overall kernel is simply the product of the two individual kernels. 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. 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, an la_i-la. Though not shown here, the predicted

[10] 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).

values will look similar to the sum of the trend and signal components of the spatial AR models shown in Figure 3 and 4.

In practice, researchers in urban economics, regional science, and geography have used a variant of LWR that has come to be known as "geographically weighted regression" (GWR). In Figures 1-4, the base model takes the form $y_i = \beta_0 + \beta_1 x_i + u_i$. My emphasis so far has been on finding a simple relationship between y and x, but we might also consider a variant of the model in which the coefficients β_0 and β_1 vary spatially, i.e., they are 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" (CPAR) because given the values of lo and la, the model is a simple parametric function of x. The CPAR model is considered in detail by Cleveland, Grosse, and Shyu (1992) and Cleveland (1994), while still more general versions are analyzed in Hastie and Hibshirani (1993). The spatial version of the model was first used by McMillen (2006), although the GWR name was coined by Brunsdon, Charlton, and Fotheringham (2006).

In practice, the kernel function for the spatial version of the CPAR model is more often based on the simple distance between two points rather than the underlying geographic coordinates. Thus, if lo and la are the target coordinates and lo_i and la_i are the coordinates for observation i, then the distance between the target point and observation i is simply d_i. A univariate kernel can then be used, i.e., $K\left(\frac{d_i}{h}\right)$, and the estimated value of y at the target point is constructed using a weighted least squares regression of y on x_i-x. This approach reduces the problem to what amounts to a simple univariate LWR model, but the argument in the kernel is d_i/h rather than $(x_i-x)/h$.

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 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 will 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.

6. Multivariate Regressions with Multiple Measures of Proximity

Non-parametric estimation suffers from a "curse of dimensionality" – the variance of the estimates increases rapidly with the number of variables. In this situation, semi-parametric models become an attractive alternative to full non-parametric estimation. Semi-parametric models separate the equation into parametric and non-parametric components, i.e.,

$$Y = X\beta + f(Z) + u \qquad (7)$$

where X is a set of variables whose effects on Y are assumed to be modeled adequately using a simple parametric function and Z is a set of variables whose effects enter the equation non-parametrically. 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 semi-parametric 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 non-parametric modeling to Z alone.

Following Robinson (1988), the steps for estimating the semi-parametric model are:

1. Use a non-parametric estimation procedure – kernel regression, LWR, CPAR, etc. – to construct predicted values for Y and each variable in X using Z as explanatory variables. The residuals from these non-parametric regressions are e_y and e_x.
2. Regress e_y on e_x using standard linear regression procedures. The coefficients on e_x are the estimated values of β, and the usual standard errors from this regression are appropriate for $\hat{\beta}$. The predicted value for the parametric part of the regression are $X\hat{\beta}$ ($not\ e_X\hat{\beta}$).
3. Use a non-parametric procedure to estimate f(Z); the dependent variable is $Y - X\hat{\beta}$ and the explanatory variables are Z.
4. The predicted value of the dependent variable is $\hat{Y} = X\hat{\beta} + \hat{f}(Z)$.

Thus, the semi-parametric estimator involves no additional programming effort beyond that required to construct the simple non-parametric models discussed in previous sections.

It is also possible to use flexible series expansions in place of the semi-parametric model. Generalizations of the fourier and spline approaches are available for two variables. Indeed, these approaches are probably preferable to a semi-parametric approach if the goal is simply to allow for nonlinear effects in a single variable such as distance from the CBD. However, the semi-parametric 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.

Table 3 presents fixed effects estimates and Table 4 presents semi-parametric results for a representative hedonic price function. The data set includes all sales of small residential homes (six units or fewer) in Chicago for 1990-1991 and 1993-1999 (data for 1992 are unavailable). With 82,807 observations, the spatial AR and spatial error models are not even feasible to estimate. In addition to standard housing characteristics, the regression includes controls for several variables that are explicitly spatial: proximity to an elevated train line (the "EL") or other rail line, distance from the Chicago CBD, distance from an EL stop, and distance from Lake Michigan.[11] 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.[12]

Table 3 presents descriptive statistics, a base linear model, and two specifications with spatial 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. The fixed effects estimator becomes a very blunt instrument when only a few observations are in a group. Since identification comes from the deviations of 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

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

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

for variables like distance from the CBD when the number of fixed effects increases from 76 to 764 – precisely what would be expected from a variable that has very little variation within an area as small as a census tract.

Table 4 includes comparable semi-parametric specifications, substituting smooth spatial trends in the geographic coordinates for the sharp discontinuities at zone boundaries implied by the fixed effects specification.[13] Although the results for the housing structural characteristics are fairly stable as the window sizes increase from 10% to 70%, the coefficients for the explicitly spatial variables vary a great deal. In general, the estimates reach much higher levels of statistical significance as the window size increases. While the two variables of most interest – distance from an EL stop and the number of murders – are highly significant across all window sizes, the coefficients vary enough to cause significant changes in any cost-benefit calculations that might follow from the estimates. Qualitatively, the results of both the fixed effects and semi-parametric specifications are similar: proximity to EL stops and a high incidence of crime are capitalized into home values as expected. Unfortunately, the magnitudes of these effects are quite sensitive to reasonable variations in the specification.

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 or proximity to Lake Michigan. 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

[13] I use a tri-cube product kernel and a LWR specification for the nonparametric portions of the estimation procedure. The nonparametric portion of the model includes normalized values of the latitudes and longitudes. I use the excellent LOCFIT library in R to do the calculations. This procedure includes a sophisticated interpolation procedure that produces extremely fast predictions at each data point for very large data sets. It also produces calculations of the degrees of freedoms used in estimation and the GCV and CV values.

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 semi-parametric 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.

Though either fixed effects or semi-parametric specification can help guide the model specification, they have significant differences. First, it is important to recognize that even the small 10% window can use significantly fewer degrees of freedom to control for spatial trends than a fixed effects specification. In the example here, the number of degrees of freedom used to estimate f(z) ranges from about 7 for the 70% window to 40 for the 10% window. In contrast, the two fixed effects specifications use 76 and 764 degrees of freedom to accomplish the same objective. Second, the fixed effects model produces discontinuities at boundaries between zones, whereas the usual semi-parametric specification leads to a smooth surface. If there truly are separate effects shared by all observations in a zone, then the fixed effects model is the correct one. But if the goal is to account for broad spatial effects using a fairly arbitrary set of fixed effects, a non-parametric approach is likely to be better.

7. Extensions and Alternatives

Although the literature on non-parametric and semi-parametric spatial data analysis is extensive, important work remains to be done. Perhaps most importantly, research is needed to establish rules for choosing bandwidths and window sizes when the objective is to estimate marginal effects rather than prediction. Much of the concern about highly variable coefficients in CPAR models is a result of choosing small bandwidths based on rules designed to produce good predicted values of the dependent variable. In general, more attention needs to be paid to methods for hypothesis testing in non-parametric models.

Work is also needed in generalizing existing non-parametric methods to spatial data. There is no reason that the fixed effects estimator – which produces discontinuities at boundaries between zones – should be the primary alternative to non-parametric estimation with smooth spatial kernels. It is possible to use left and right sided kernels (e.g., using all points to the west or east of a target points) rather than kernels that are centered on the target location. Many years ago, McDonald and Owen (1986) proposed a "split linear fit" procedure that uses a weighted average of right, left, and centered windows (along with variations in the bandwidth) to estimate a model that allows the data to determine the appropriate degree of smoothness. To my knowledge, my first attempt at non-parametric modeling remains the only application of this sort of estimator to spatial data (McMillen, 1994). This sort of non-parametric estimator has the potential to combine the best features of non-parametric estimation and fixed effects models.

Although spline functions have a long history in urban economics, series expansions are still not commonly used in spatial model. This too is an area where more familiarity with the statistical literature could produce large payoffs for spatial modeling.

Spatial data analysis can be simpler when panel data are available or when geographic zones are discrete. For example, consider Black's (1999) influential paper using house prices to measure the value of 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). Similar estimation strategies have been used by such authors as Cunningham (2007) to analyze the effects of growth controls and Greenstone and Gallagher (2008) to analyze the benefits Superfund-sponsored cleanups of hazardous waste sites. Although the 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.[14] More generally, a potential problem with the approach is that it requires that spatial effects be confined to a simple discrete change at zone boundaries. Spatial variation within zones can severely bias the results.

With panel data, it may be possible to isolate the effect of changes in a critical variable, e.g., the effect of new transit lines on house prices (Baum-Snow and Kahn, 2000; Gibbons and

[14] Greenstone and Gallagher's (2008) approach is somewhat different from Black's in that they rely on a regression discontinuity approach, comparing "housing market outcomes in the areas surrounding the first 400 hazardous waste sites chosen for Superfund cleanups to the areas surrounding the 290 sites that narrowly missed qualifying for these cleanups" (Greenstone and Gallagher, 2008, p. 951). But the issue is similar in that the approach is based on an assumption that the areas that just missed Superfund designation are not systematically different from qualifiers.

Machin, 2005; McMillen and McDonald, 2004). Indeed, this idea is the central idea of repeat sales models (Case and Shiller, 1989). But again the approach may not be entirely effective if there is spatial variation in missing variables that also influence house price changes. The standard repeat sales approach hinges on the assumption that the effects of other variables – both the values of the variables and their coefficients – are not changing over time. In general, models taking advantage of "natural experiments" to analyze the effects of a change in a variable rely heavily on an assumption that no other important changes took place, or that the changes are taken into account adequately using various control variables.

8. 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 data sets 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.

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 non-parametric 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.

References

Abelson, Peter 1997. "House and Land Prices in Sydney from 1931 to 1989," *Urban Studies,* 34, 1381-1400.

Ahlfeldt, Gabriel M. and Nicolai 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, William 1964. *Location and Land Use*. Cambridge, MA: Harvard University Press.

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

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

Atack, Jeremy and Robert 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.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Econom*ics, 114, 577-599.

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

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

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

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

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

Brueckner, Jan K. 2006. "Strategic Interaction Among Governments," in Richard J. Arnott and Daniel P. McMillen (eds.), *A Companion to Urban Economics*. Malden, MA: Blackwell, 332-347.

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

Brunsdon, Chris, A. Stewart Fotheringham, and Martin. Charlton 1996. "Geographically Weighted Regression," *Geographical Analysis*, 28, 281-298.

Cameron, Trudy A. 2005. "Directional heterogeneity in Distance Profiles in Hedonic Property Value Models," *Journal of Environmental Economics and Management*, 51, 26-45.

Case, Anne C., Harvey S. Rosen, and James R. Hines 1993. "Budget Spillovers and Fiscal Policy Interdependence: Evidence From the States," *Journal of Public Economics*, 52, 285–307.

Case, Karl E. and Robert J. Shiller 1989. "The Efficiency of the Market for Single-Family Homes," *American Economic Review*, 79, 125-137.

Clark, Colin 1951. "Urban Population Densities," *Journal of the Royal Statistical Association Series A*, 114, 490-496.

Cleveland, William 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, 21-36.

Cleveland, William S. and Susan J. Devlin 1988. "Locally Weighted Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596-610.

Cleveland, William S., E. H. Grosse, and W. M. Shyu 1991. "Local Regression Models," in J. M. Chambers and T. J. Hastie (eds.), *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks/Cole, 309-376.

Coulson, N. Edward 1991. "Really Useful Tests of the Monocentric City Model," *Land Economics*, 67, 299-307.

Cunningham, Christopher R. 2007. "Growth Controls, Real Options, and Land Development," *Review of Economics and Statistics*, 89, 343-358.

Fredriksson, Per G. and Daniel L. Millimet 2002. "Strategic Interaction and the Determinants of Environmental Policy Across U.S. States," *Journal of Urban Economics*, 51, 101–122.

Gallant, Ronald 1981. "On the Bias in Flexible functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211-245.

Gallant, Ronald 1982. "Unbiased Determination of Production Technologies," *Journal of Econometrics*, 20, 285-323.

Gibbons, Steve 2004. "The Costs of Urban Property Crime," *Economic Journal*, 114, 441-463.

Gibbons, Stephen and Stephen Machin, 2005. “Valuing Rail Access using Transport Innovations,” *Journal of Urban Economics*, 57, 148-169.

Greenstone, Michael and Justin Gallagher 2008. “Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program,” *Quarterly Journal of Economics*, 123, 951-1003.

Härdle, Wolfgang and Oliver B. Linton 1994. “Applied Nonparametric Methods,” in Robert F. Engle and Daniel L. McFadden (eds.), *Handbook of Econometrics, Volume 4*. New York: North-Holland, 2295-2339.

Hastie, Trevor and Robert Tibshirani 1993. “Varying-Coefficient Models,” *Journal of the Royal Statistical Society, Series B,* 55, 757-796.

Ihlanfeldt, Keith 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, Qi and Jeffrey S. Racine 2007. *Nonparametric Econometrics*. Princeton, NJ: Princeton University Press.

Loader, Clive 1999. *Local Regression and Likelihood*. New York: Springer.

McDonald, John A. and Art B. Owen 1986. “Smoothing with Split Linear Fits,” *Technometrics*, 28, 195-208.

McDonald, John F. 1989. "Econometric Studies of Urban Population Density: A Survey," *Journal of Urban Economics*, 26, 361-385.

McMillen, Daniel P. 1994. “Vintage Growth and Population Density: An Empirical Investigation,” *Journal of Urban Economics*, 36, 333-352.

McMillen, Daniel P. 1996. “One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach,” *Journal of Urban Economics*, 40, 100-124.

McMillen, Daniel P. 2006. “Testing for Monocentricity,” in Richard J. Arnott and Daniel P. McMillen (eds.), *A Companion to Urban Economics*. Malden, MA: Blackwell, 128-140.

McMillen, Daniel P. and Jonathan Dombrow 2001. “A Flexible Fourier Approach to Repeat Sales Price Indexes,” *Real Estate Economics*, 29, 207-225.

McMillen, Daniel P. and John F. McDonald 2004. “Reaction of House Prices to a New Rapid Transit Line: Chicago’s Midway Line,” *Real EstateEconomics*, 32, 463-486.

Meese, Richard and Nancy 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, Daniel L. and Vasudha Rangaprasad 2007. “Strategic Competition Amongst Public Schools,” Regional Science and Urban Economics, 37, 199-210.

Mills, Edwin S. 1969. “The Value of Urban Land,” in Harvey Perloff (ed.), *The Quality of the Urban Environment*. Baltimore: Resources for the Future, Inc., 231-253.

Mills, Edwin S. 1972. Studies *in the Structure of the Urban Economy*. Baltimore: Johns Hopkins Press.

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

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

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

Pope, Davin G. and Jaren C. Pope 2009. “Crime and Property Values: Evidence from the 1990’s Crime Drop,” Blacksburg, VA: Virginia Tech University.

Robinson, Paul M. 1988. ‘‘Root-N-Consistent Semiparametric Regression,’’ *Econometrica*, 56, 931–954.

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

Schwartz, Amy E., Scott Susin, and Ioan Voicu 2003. "Has Falling Crime Driven New York City's Real Estate Boom?" *Journal of Housing Research*, 14, 101-135.

Smith, Fred H. 2003. “Historical Evidence on the Monocentric Urban Model: A Case Study of Cleveland, 1915-1980,” *Applied Economics Letters*, 10, 729-731.

Suits, Donald B., Andrew Mason, and Louis Chan 1978. “Spline Functions Fitted by Standard Regression Models,” *Review of Economics and Statistics*, 60, 132-139.

Thorsnes, Paul, Robert Alexander, and Bruce McLennan 2009. “Low-Income Housing Built in High-Amenity Area: Long Run Housing-Market Effects of Exogenous Amenities,” Dunedin: University of Otago.

Thorsnes, Paul and John W. Reifel 2007. “Tiebout Dynamics: Neighborhood Response to a Central-City/Suburban House-Price Differential,” *Journal of Regional Science*, 47, 693-719.

Yatchew, Adonois 1998. “Nonparametric Regression Techniques in Economics,” *Journal of Economic Literature*, 36, 669-721

Table 1

The Log of Floor Area Ratios: Series Expansions

Variable	Base Linear Model	Spline	Variable	Fourier
Constant	-0.078 (4.017)	0.626 (12.634)	Constant	3.827 (3.662)
x = Distance to City Center	-0.072 (46.187)	-0.223 (12.987)	$z = 2\pi x/33.620$	-3.740 (3.775)
x^2		0.006 (2.395)	z^2	0.518 (3.295)
x^3		-7.06e-07 (0.015)	sin(z)	-0.488 (6.594)
$(x-16.811)^3$ if x>16.811, 0 otherwise		-2.38e-04 (2.144)	cos(z)	-1.609 (2.537)
			sin(2z)	-0.155 (3.884)
			cos(2z)	-0.424 (2.656)
			sin(3z)	-0.112 (3.721)
			cos(3z)	-0.187 (2.714)
			sin(4z)	-0.036 (1.514)
			cos(4z)	-0.089 (2.483)
			sin(5z)	-0.017 (0.926)
			cos(5z)	-0.046 (2.275)
R^2	0.618	0.780		0.786

Notes. Absolute t-values are in parentheses below the estimated coefficients. The number of observations is 1322.

Table 2

Spatial AR Models for the Log of Floor Area Ratios

	Base Linear Model	Spline
Constant	-0.007 (0.572)	0.172 (4.220)
x = Distance to City Center	-0.013 (7.781)	-0.058 (4.098)
x^2		7.63e-04 (0.594)
x^3		2.88e-05 (0.793)
$(x-16.811)^3$ if $x>16.811$, 0 otherwise		-1.65e-04 (1.923)
WY	0.833 (45.140)	0.709 (27.459)

Notes. Absolute z-values are in parentheses below the estimated coefficients. The number of observations is 1322.

Table 3
House Price Regressions with Fixed Effects for Location

Variable	Mean (std. dev.)	Base Model	Community Area Fixed Effects (76 areas)	Census Tract Fixed Effects (764 tracts)
Log of Building Area	7.057 (0.507)	0.423 (66.767)	0.363 (72.614)	0.325 (69.564)
Log of Land Area	8.259 (0.295)	0.211 (46.122)	0.234 (62.169)	0.216 (58.818)
Age	63.451 (0.325)	-0.003 (50.749)	-0.003 (52.143)	-0.003 (46.940)
Number of Bedrooms	2.846 (25.001)	0.003 (1.300)	0.019 (11.379)	0.023 (15.234)
2+ Stories	0.341 (0.773)	0.022 (7.008)	-0.054 (22.387)	-0.056 (24.515)
Brick	0.621 (0.150)	-0.017 (5.379)	0.039 (15.366)	0.018 (7.524)
Basement	0.773 (0.485)	0.028 (7.596)	-0.004 (1.348)	0.004 (1.672)
Attic	0.459 (0.419)	-0.020 (7.194)	-0.007 (3.174)	-0.012 (6.273)
Central Air Conditioning	0.189 (0.498)	0.094 (28.114)	0.018 (6.778)	0.008 (3.472)
1-Car Garage	0.320 (0.391)	0.044 (13.464)	0.046 (17.941)	0.040 (17.253)
2+ Car Garage	0.448 (0.466)	0.066 (20.796)	0.070 (28.141)	0.065 (28.592)
Fireplace	0.091 (0.497)	0.168 (37.464)	0.093 (25.672)	0.059 (17.380)
Within ¼ Mile of El Line	0.094 (0.288)	0.165 (36.347)	0.057 (14.643)	0.010 (2.122)
Within ¼ Mile of Rail Line	0.442 (0.292)	-0.076 (29.991)	-0.020 (9.444)	-0.021 (8.549)
Distance from CBD	8.938 (0.497)	-0.053 (76.887)	0.017 (7.221)	0.005 (0.804)
Distance from El Stop	1.633 (2.640)	0.070 (45.409)	0.045 (16.202)	0.013 (2.465)
Distance from Lake Michigan	5.507 (1.138)	2.44e-05 (0.040)	-0.004 (1.726)	0.027 (4.651)
Number of Murders in Census Tract during Year	27.820 (2.416)	-0.012 (172.885)	-0.001 (10.769)	-0.001 (6.624)
R^2		0.552	0.737	0.788

Notes. Absolute z-values are in parentheses below the estimated coefficients. The number of observations is 82807. The regressions also include intercepts and eight variables indicating the year of sale.

Table 4
Semi-Parametric House Price Regressions

Variable	10% Window	30% Window	50% Window	70% Window
Log of Building Area	0.366 (73.437)	0.372 (71.417)	0.381 (68.868)	0.378 (66.815)
Log of Land Area	0.237 (63.333)	0.247 (64.111)	0.234 (57.676)	0.217 (52.571)
Age	-0.003 (50.636)	-0.003 (54.757)	-0.003 (57.752)	-0.003 (58.303)
Number of Bedrooms	0.016 (9.945)	0.015 (8.462)	0.010 (5.598)	0.009 (4.907)
2+ Stories	-0.053 (21.943)	-0.043 (16.766)	-0.026 (9.737)	-0.020 (7.105)
Brick	0.027 (10.935)	0.028 (10.852)	0.022 (7.828)	0.020 (7.128)
Basement	-0.003 (1.064)	0.004 (1.438)	0.006 (1.813)	0.003 (0.985)
Attic	-0.009 (4.451)	-0.011 (5.052)	-0.009 (3.874)	-0.011 (4.429)
Central Air Conditioning	0.022 (8.265)	0.027 (9.598)	0.035 (11.744)	0.044 (14.661)
1-Car Garage	0.048 (18.756)	0.048 (18.025)	0.051 (17.968)	0.052 (17.721)
2+ Car Garage	0.072 (29.073)	0.073 (27.815)	0.078 (28.117)	0.080 (28.124)
Fireplace	0.094 (26.138)	0.123 (32.826)	0.148 (37.512)	0.158 (39.369)
Within ¼ Mile of El Line	0.032 (8.279)	0.041 (10.474)	0.079 (19.465)	0.102 (24.920)
Within ¼ Mile of Rail Line	-0.035 (15.970)	-0.026 (11.893)	-0.023 (10.182)	-0.024 (10.455)
Distance from CBD	0.005 (0.571)	0.012 (3.010)	-0.011 (4.417)	-0.014 (9.226)
Distance from El Stop	0.010 (2.485)	0.037 (17.011)	0.058 (33.619)	0.062 (41.236)
Distance from Lake Michigan	-0.003 (0.205)	-0.110 (12.516)	0.028 (3.574)	0.076 (11.141)
Number of Murders in Census Tract during Sale Year	-0.002 (17.468)	-0.003 (32.326)	-0.005 (60.435)	-0.006 (74.081)
Degrees of Freedom Used in Non-parametric Portion	39.698	15.327	8.982	6.806

Notes. Absolute z-values are in parentheses below the estimated coefficients. The number of observations is 82807. The regressions also include eight variables indicating the year of sale.

Figure 1

Data Plot for Log of Average Floor Area Ratios in Cook County Census Tracts

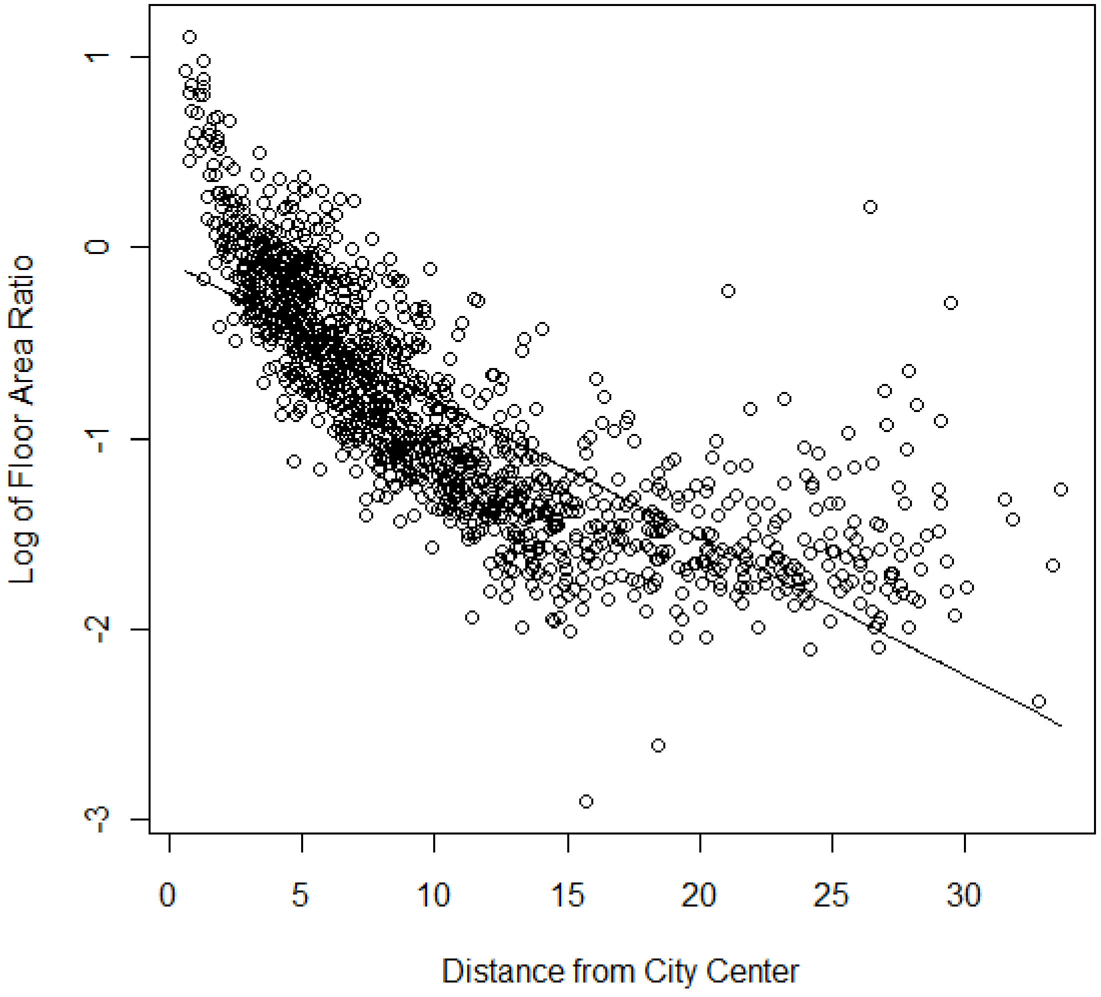


Figure 2

Estimated Linear, Spline, Fourier, and LWR Functions

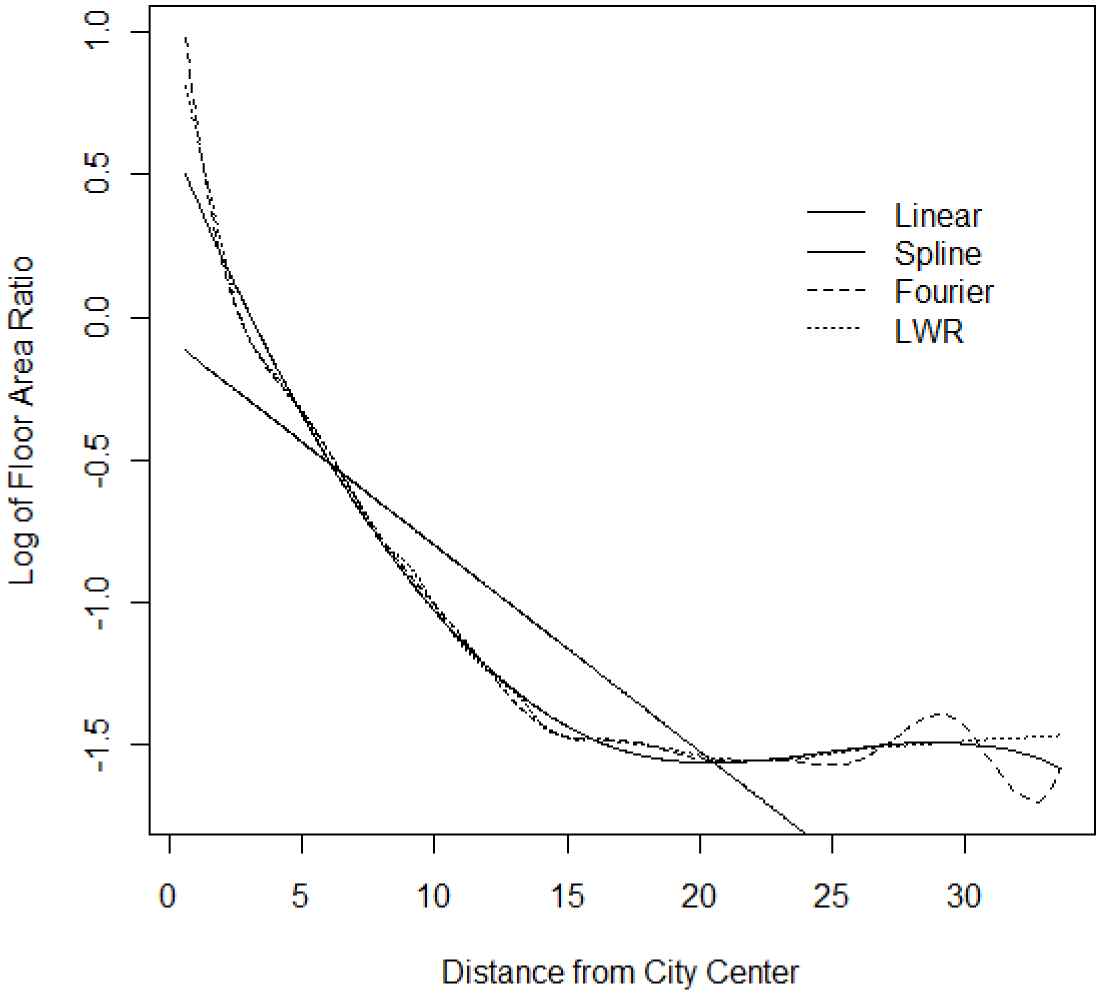


Figure 3

Comparison of Spatial AR with Linear Base Function and LWR Predictions

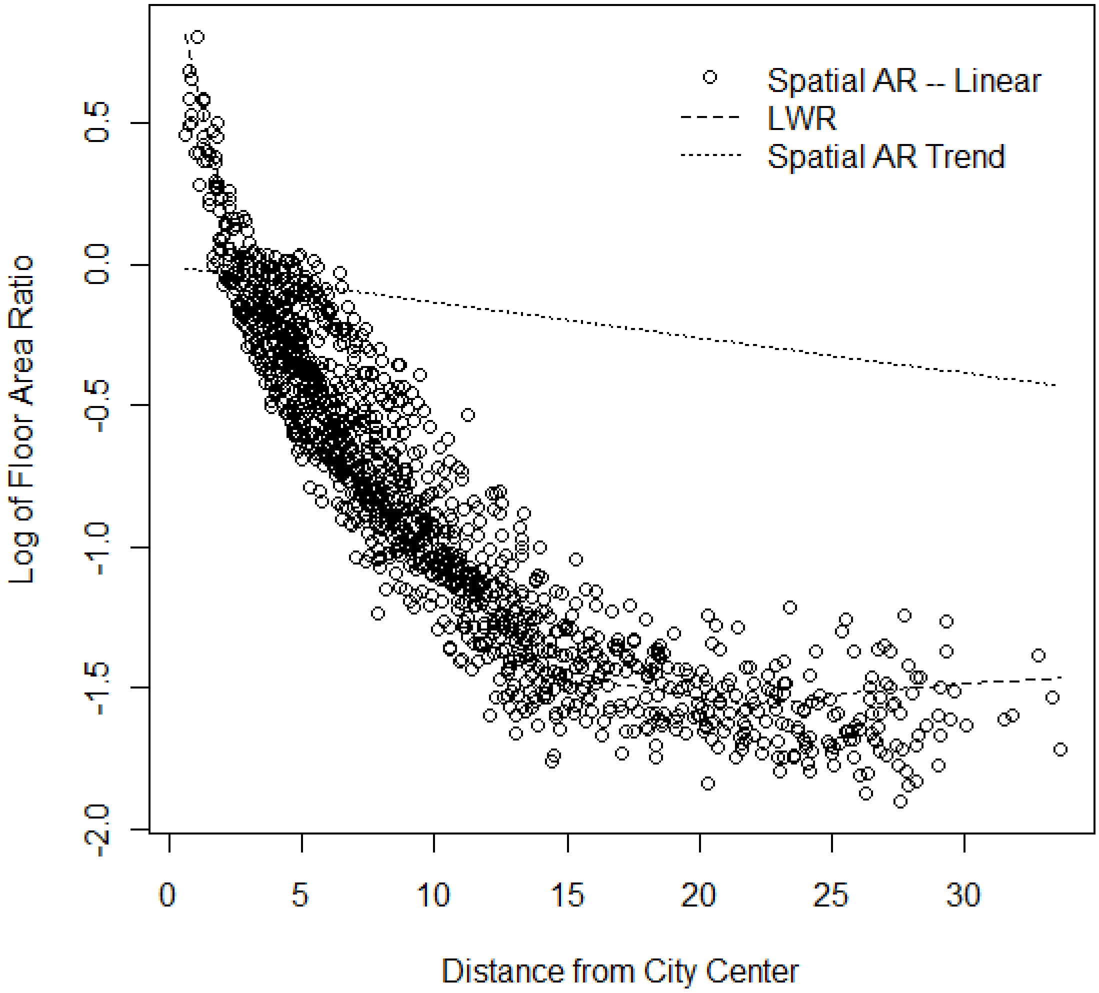


Figure 4

Comparison of Spatial AR with Spline Base Function and LWR Predictions

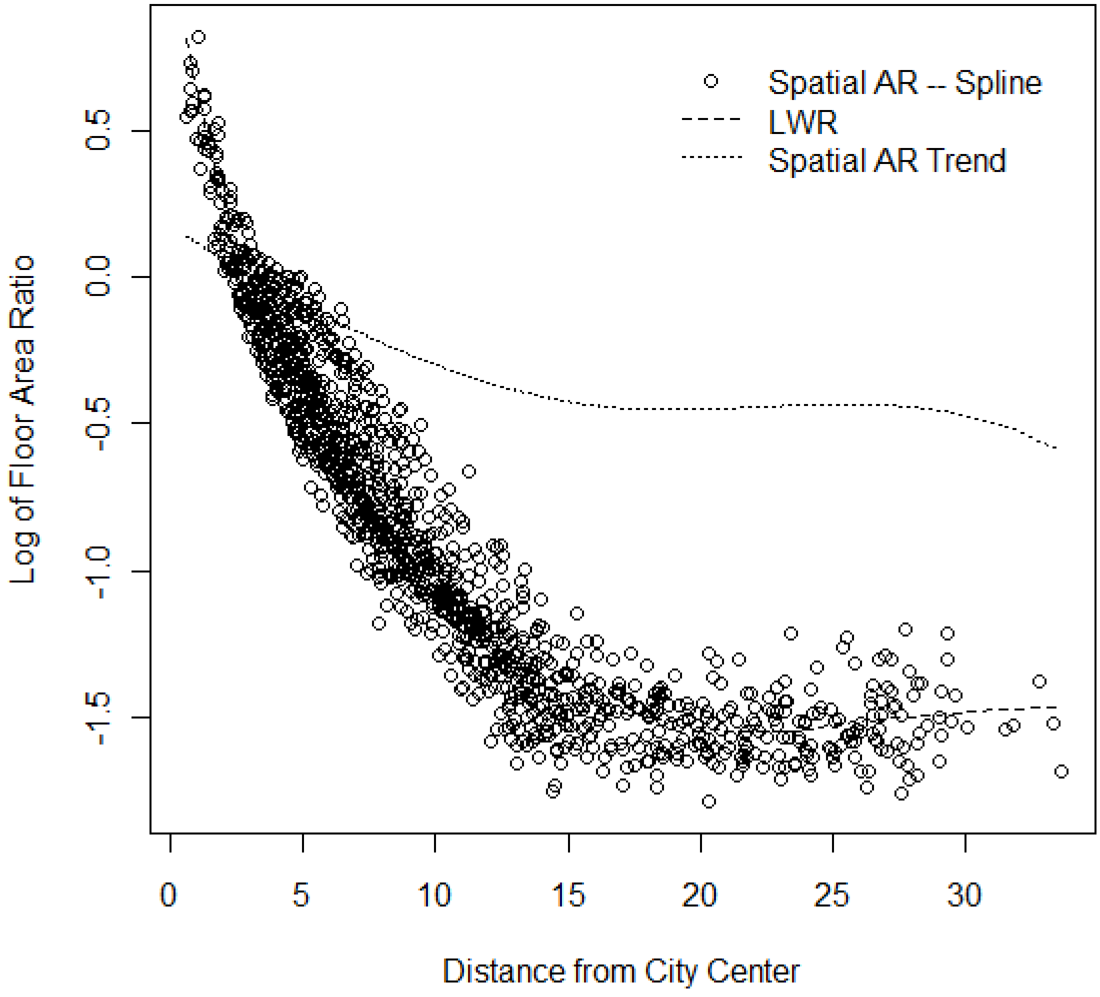
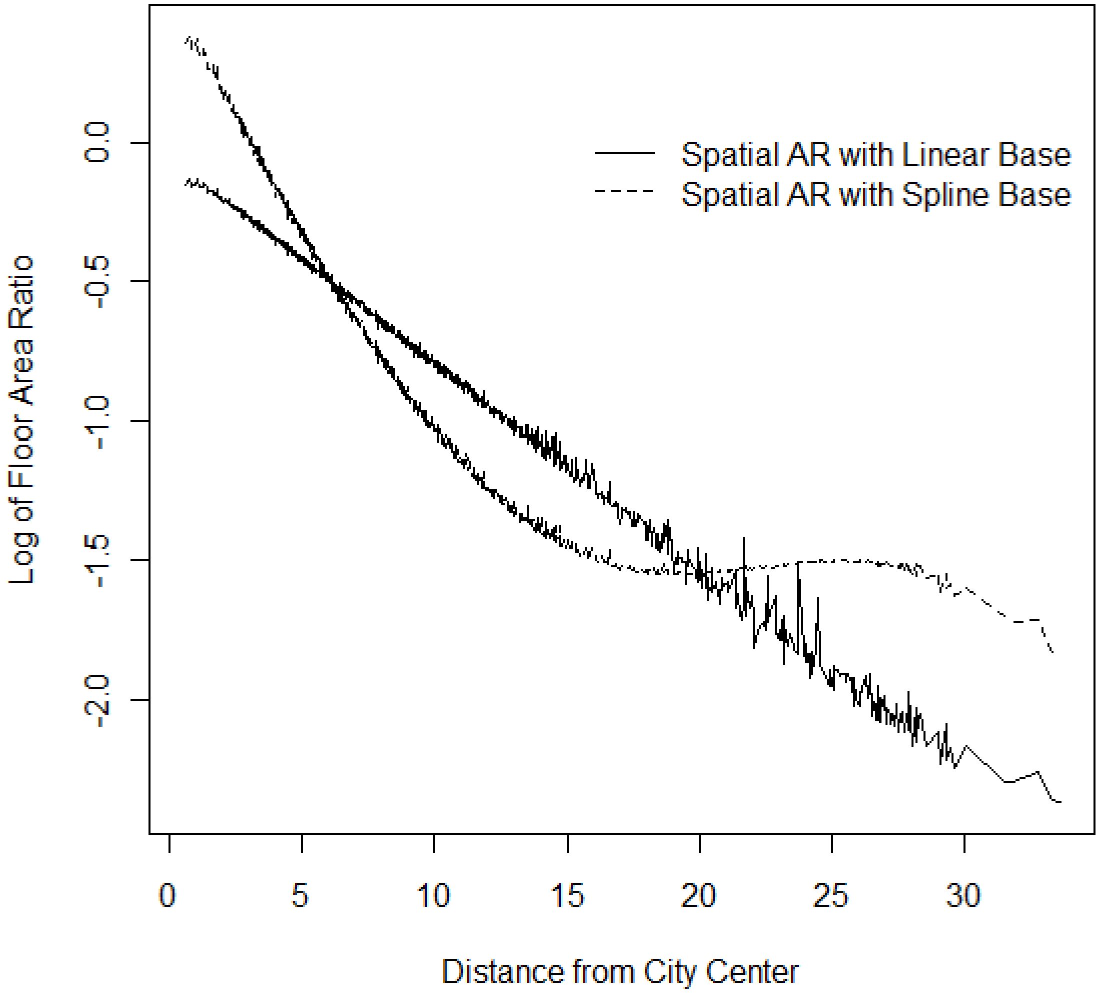


Figure 5

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


Spatial AR with Linear Base
Spatial AR with Spline Base
Log of Floor Area Ratio
Distance from City Center