Estimation, Interpretation, and Hypothesis Testing for Nonparametric Hedonic House Price Functions

Daniel P. McMillen
Institute of Government and Public Affairs
Department of Economics
University of Illinois at Chicago
601 S. Morgan St.
Chicago, IL 60607
Phone: (312) 413-2100
Fax: (312) 996-3344
Email: mcmillen@uic.edu

Christian Redfearn
School of Policy, Planning and Development
University of Southern California
Los Angeles, CA 90089-0626
Phone: (213) 821-1364
Fax: (213) 740-0001
Email: redfearn@usc.edu

October 2, 2007

Abstract

Despite well-documented shortcomings, hedonic and repeat sales estimators remain the most widely used methods for constructing quality controlled house price indexes and for assessing housing attribute capitalization into dwelling prices. Nonparametric estimators overcome many of the problems associated with these approaches by controlling for misspecified spatial effects while using highly flexible forms. Despite these advantages, nonparametric procedures are still not used extensively for data analysis due to perceived difficulties associated with estimation and hypothesis testing. We demonstrate that nonparametric estimation is both feasible for large data sets with many explanatory variables and offers significant advantages in terms of the information content of the estimates. These features are demonstrated in an application of valuing capitalization of access to a rapid transit line into surrounding dwelling prices.

1. Introduction

Hedonic and repeat sales estimators remain the dominant tools of researchers interested in either aggregate house price trends or the capitalization of housing characteristics into dwelling prices. Indeed, to a great extent what we understand of the behavior of housing markets, the value of public goods, the value of access, and the values of individual physical and structural dwelling attributes is derived from various specifications of hedonic models. Aggregate house price indexes based on hedonic analyses are relatively rare due to data requirements, but repeat sales indexes -- which are derived from a hedonic specification -- are the most widely quality controlled indexes for national and metropolitan housing markets (OFHEO, S&P/Case-Shiller). The widespread adoption of these estimators as standard practice belies some significant drawbacks to these approaches – both theoretical and empirical. Though these have been documented extensively, the continued use of these estimators reflects a perceived lack of alternatives. In this paper, we argue that nonparametric approaches offer both viable and attractive alternatives for use in housing market analyses.

Nonparametric models offer significant advantages for hedonic price function estimation. For example, functional form flexibility is a common feature of all nonparametric procedures, but one that is largely ignored in the practice of hedonic analysis. In addition, nonparametric procedures are readily adaptable to the special requirements of spatial data sets, allowing coefficients to vary by submarket over both space and time. Unlike standard parametric spatial models, this combination of functional form flexibility and spatially varying coefficients helps to reduce spatial autocorrelation without imposing arbitrary contiguity matrices or distributional

assumptions on the data. Despite these advantages, nonparametric procedures are still not used routinely in analyzing housing data sets. Several factors account for this apparent unpopularity, including perceptions that nonparametric estimation procedures (1) are difficult to implement, (2) produce results that are difficult to interpret, (3) are wasteful of degrees of freedom, and (4) are more appropriate for prediction than for hypothesis testing.

In this paper, we use a combination of Monte Carlo techniques and a representative housing data set to illustrate how nonparametric estimation procedures can readily be used in formal hypothesis testing and as an informal means of checking a model specification. Moreover, we show that standard hedonic analysis is nested within the more general locally-weighted regression (LWR) framework and that the inherent flexibility of LWR allows for the capture of a much richer set of housing market dynamics. That is, where the standard hedonic approach generally requires a single implicit price for housing attributes and local amenities, locally-weighted regression provides a set of prices that can in turn be used to examine the fundamentals of differential pricing across submarkets – for example, the value of commuter rail access may be high in neighborhoods where residents commute to work in the CBD, and it may be low in places where people incur the noise and nuisance of the train line while commuting elsewhere. The possibility of differential pricing is precluded in standard applications of hedonic analysis, and with it is the opportunity to examine local housing

market fundamentals relevant to both policy and our basic understanding of housing markets.[1]

Our focus is on locally weighted regression (LWR), which under various pseudonyms has become the most commonly employed nonparametric procedure for analyzing spatial data. We show that kernel regression and the conditionally parametric model are special cases of LWR, while the estimator that has come to be known as "geographically weighted regression" is a special case of the special case. Using Monte Carlo procedures, we show that LWR uses far fewer degrees of freedom than fully nonparametric estimation while achieving impressive degrees of predictive accuracy. Following the seminal work of Cleveland and Devlin (1988), we show how LWR can readily be used to conduct formal hypothesis tests for large data sets.

Using a representative data set of single-family home sales in Chicago for 2000, we show how a combination of nonparametric and semi-parametric procedures can be used to assist in the specification of a hedonic house price function. Our case study focuses on the valuation of access to an important local amenity, Chicago's elevated rapid transit line (the "EL") – though any housing attribute could be analyzed analogously. While the magnitude of the coefficient on distance to the nearest EL stop is reduced significantly after taking into account local spatial effects by either LWR or through census tract fixed effects, LWR uses far fewer degrees of freedom than the fixed-effect method. Though counterintuitive, the local regression's imposition of smoothness greatly reduces the information consumed in each regression. Maps of the LWR (reported below) coefficients reveal that proximity to the EL is not valued equally

[1] Of course, standard hedonic analysis can be extended to allow for spatially-varying implicit prices using interactions with geographical dummy variables. However, it is generally not feasible to estimate models with a large number of interactions due to a lack of degrees of freedom. This is discussed in detail below.

throughout the city. In higher-income neighborhoods on the north and southwest sides of the city, proximity to EL stops is highly valued for the access it provides to Chicago's central business district. In lower-income neighborhoods on the west and south sides of the city, proximity to the EL appears to be a disamenity.[2] In these neighborhoods, the EL is lined by vacant lots and failing businesses, which leads to a 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. Moreover, in a traditional specification the underlying spatial variation in pricing would not be available for subsequent analysis in which the determinants of the various values of access to the EL could be studied.

The implications of such misspecification are significant because so much of what we understand about the market valuation of spatially varying amenities and disamenities is based on capitalization studies that use standard hedonic analysis. We show that typical approaches to assessing the value of access to the EL yield a population average that is not readily interpretable for use in guiding policy. That is, standard hedonic analysis finds a significant relationship between access to the EL and dwelling prices, but this average effect masks widely-varying local valuations that appear to have ready explanation via standard microeconomic theory.

The paper is organized as follows. In Section 2, we lay out the mechanics of nonparametric analysis and contrast three particular methods with standard hedonic analysis: locally-weighted regression, kernel regression, and conditionally parametric

[2] Bowes and Ihlanfeldt (2001) estimate the components of access, finding that distance to rail stations contains both amenities (lower commute times, access to retail) and disamenities (congestion, crime). This disaggregation of the coefficient on distance to stations is not the focus of this paper, but offers one explanation as to the economic forces that would generate spatially-varying coefficients.

regression. In Section 3, we discuss hypothesis testing and degrees of freedom calculations in these contexts. In Sections 4 and 5, we illustrate the statistical characteristics of the three nonparametric approaches using both Monte Carlo simulations and representative data from Chicago. Section 6 concludes.

2. Nonparametric Estimation Procedures

The standard parametric specification of a hedonic housing price function is $y_i = \beta' X_i + u_i$, where y_i is the observed sales price for observation i (or more commonly, the natural logarithm of sales price), X_i is a vector of explanatory variables, and u_i is an error term. The vector X_i is typically divided into variables representing characteristics of the home itself, S_i, and characteristics of the location, N_i. Examples of the former include square footage and lot size, while examples of locational characteristics include distance from the city center or indicators of neighborhoods and school districts.

The classical statistical assumption of a known model specification is always violated in practice. At least since Rosen's (1974) seminal paper developing the theory of hedonic price functions, empirical researchers have recognized that hedonic house price functions are likely to be nonlinear in structural characteristics, and there is no reason to expect prices to be linear in continuous measures of locational variables such as distance from the city center. Due to a lack of arbitrage opportunities for a fixed and immalleable housing stock, the housing characteristics' contributions to price may vary over space – i.e., there may be interactions between the S and N variables.[3] This

[3] The existence of submarkets and their causes are well-reported; see, for example Goodman and Thibodeau (2007, 1997) and their bibliographies for others. One feature of submarkets is variation in implicit prices across housing market segments – something not generally allowed for in common applications of hedonic analysis.

combination of nonlinearity and spatial heterogeneity suggests that the standard parametric model is a particularly convenient and simple specification of the more general model $y_i = f(X_i) + u_i$ or $y_i = f(S_i, N_i) + u_i$.

In the past, the Box-Cox transformation was employed routinely in hedonic price estimation (e.g., Pollakowsi and Halvorsen (1981)). Though generally lacking theoretical basis, polynomial terms or spline functions are still used commonly to approximate the unknown function with a parametric equation. These steps toward statistical flexibility lead naturally to nonparametric procedures which have begun to be used in various forms. The most commonly employed nonparametric procedures are (1) locally weighed regression, (2) kernel regression, and (3) conditionally parametric regression (CPAR), which includes a special case that is sometimes called geometrically weighted regression.

Each of these procedures fits individual regressions targeted to specific points, with more weight placed on observations that are closer to the target. "Closer" can be defined narrowly in terms of geographic distance, or in terms of more general measures of distance among the full set of explanatory variables. For example, the distance function could be based on search theory – arguing that dwellings that share attributes are likely to be bid on by the same set of agents and therefore have similar attribute pricing. In this case, the distance function could be defined across overall size, number of bedrooms, and location: e.g., in the regression at the observation of a 2000 square foot, three bedroom house in Elmhurst, other sales of 2000 square foot homes with 3 bedrooms nearby would get more weight than sales that were in other parts of the Chicago area (physical "closeness") or were smaller or larger (attribute "closeness"). LWR is the most general of the three procedures, with the kernel regression and CPAR being special cases.

Locally Weighted Regression

Urban applications of the LWR estimator are derived from the seminal paper by Cleveland and Devlin (1988). The first direct applications of the procedure to urban-related issues are Stock (1991) and Meese and Wallace (1991). Let the target for the nonparametric estimator be a home with structural and locational characteristics given by the vector X. The LWR estimator is derived by minimizing equation (1) with respect to α and β:

$$\sum_{i=1}^{n}\left(y_i-\alpha-\beta'\left(X_i-X\right)\right)^2 K\left(\frac{X_i-X}{h}\right) \qquad (1)$$

The kernel function $K\left(\psi\right)$ determines the weight that observation i receives in estimating the value of y at target point X. Common choices for this function when X consists of a single variable include:

Name	$K\left(\psi\right)$	Domain
Rectangular	1	$\left\|\psi\right\|<1$
Triangular	$1-\left\|\psi\right\|$	$\left\|\psi\right\|<1$
Epanechnikov	$1-\psi^2$	$\left\|\psi\right\|<1$
Bisquare	$\left(1-\psi^2\right)^2$	$\left\|\psi\right\|<1$
Tricube	$\left(1-\psi^3\right)^3$	$\left\|\psi\right\|<1$
Triweight	$\left(1-\psi^2\right)^3$	$\left\|\psi\right\|<1$
Gaussian	$e^{-.5\psi^2}$	$-\infty<\psi<\infty$

A product kernel is typically used to define $K\left(\psi\right)$ when the model includes more than one explanatory. For example, with two explanatory variables the weight function

would be $K(\psi) = K_1(\psi_1)K_2(\psi_2)$, where $\psi_1 = \left(\frac{X_{1i} - X_1}{h_1}\right)$, $\psi_2 = \left(\frac{X_{2i} - X_2}{h_2}\right)$, and the number subscript indicates the individual variable. Though it is not necessary, the same kernel function is typically used for both explanatory variables in this situation.

All of the kernels share the important feature of placing more weight on nearby observations. Partly for this reason, the choice of kernel weight function has very little effect on the results. The choice of bandwidth, *h*, is much more important. The bandwidth determines how rapidly the weights decline with distance and how many observations receive positive weight when constructing the estimate. High values of *h* produce more smoothing than low values. The bandwidth may be fixed at one value for all data points, in which case the number of observations receiving some weight in estimation varies depending on how many observations are near the target point. Alternatively, the value of *h* may vary by target point such that a fixed number of observations receive positive weight for each target. In this latter case, the bandwidth is generally referred to as the "window size" because it determines the size of the opening to observations to be included in estimation. It should be noted that ordinary least squares (OLS) is a special case within this framework: it is an application of LWR with a rectangular kernel and a bandwidth of 1; all the observations would be used and would be equally weighted so that the OLS coefficients would be recovered at each point.

There is little difference between a fixed bandwidth and a fixed window size if the observations are distributed uniformly over space. However, most spatial data sets combine regions with many observations with others where the data are sparse. A fixed bandwidth leads to excessive smoothing in areas where many observations are available for estimation near a target point, and it leads to highly variable results in areas with

sparse data. Thus, the "nearest neighbor" approach of a common window size for all target points is generally preferable to a fixed bandwidth for analyzing spatial data.[4]

The LWR model simplifies to standard weight least squares (WLS) estimation with one regression for each target point. After writing $Z_i = \begin{pmatrix} 1 & X_i \end{pmatrix}'$ and $\theta = \begin{pmatrix} \alpha & \beta \end{pmatrix}'$, the LWR estimator is:

$$\hat{\theta}(X) = \left(\sum_{i=1}^{n} K(\psi_i) Z_i Z_i' \right)^{-1} \sum_{i=1}^{n} K(\psi_i) Z_i' y_i \tag{2}$$

which is simply the vector of coefficients from a regression of $w_i y_i$ on w_i and $w_i X_i$, where $w_i = K(\psi_i)^{1/2}$. The predicted value of y at the target point is simply $Z'\hat{\theta}(X)$, i.e., the standard prediction evaluated at the target point, X. The coefficients on the explanatory variables, $\hat{\beta}(X)$, represent the estimated marginal effects at the target point. Standard errors are also easy to evaluate for $\hat{\theta}(X)$; see Pagan and Ullah (1999) or McMillen (2004a) for details.

Kernel Regression

Kernel regression is a special case of LWR in which the objective function is:

$$\sum_{i=1}^{n} (y_i - \alpha)^2 K\left(\frac{X_i - X}{h} \right) \tag{3}$$

The estimated values of y at the target point X are:

[4] Note that the window size need not be a fixed function either. Observations can be drawn by distance or by number of nearest neighbors, as described. Alternatively, jurisdictional boundaries or geographic features can be imposed. An obvious application of such rules would be school district boundaries where service flow can vary discontinuously in space.

$$\hat{y}(X) = \frac{\sum_{i=1}^{n} K(\psi_i) y_i}{\sum_{i=1}^{n} K(\psi_i)} \tag{4}$$

The marginal effects are estimated by taking the derivative of equation (4) with respect to *X*. Equation (4) can be constructed by regressing $w_i y_i$ on w_i, where $w_i = K(\psi_i)^{1/2}$. Thus, kernel regression is identical to LWR, but with only a constant term included in the WLS regression. The advantage of LWR over kernel regression is that the additional explanatory variables lead to more accurate estimates in regions with sparse data. LWR estimates are typically less variable than kernel regression estimates, allowing larger bandwidths to be used in estimation. Thus, LWR estimation is generally preferable to kernel regression for analyzing spatial data.

Conditionally Parametric Regression

Although the LWR and kernel regression models can be estimated using WLS regression techniques, the models are actually fully nonparametric. The estimators use a local linear function to approximate a function that is constrained only to be smooth and continuous. Each variable's marginal effects depend on the values taken by all other variables in the model. As will be seen in Section 4, the generality of the models comes at the cost of degrees of freedom. Since the variance of the LWR and kernel regression models increases rapidly as the number of explanatory variables increases, these estimators are typically applied only to models with relatively few explanatory variables.

The conditionally parametric model is a special case of LWR in which degrees of freedom are preserved by making the model parametric in some variables while other

variables are constrained only to have smooth and continuous marginal effects. The CPAR model applies when the vector X can be divided into portions that are fully nonparametric (X_1) and conditionally parametric (X_2), in which case the model becomes:

$$y_i = \beta_1(X_1) + \beta_2(X_1)' X_2 + u \tag{5}$$

For fixed values of X_1, this model is a standard linear equation, but each of the coefficients varies with X_1. The model is similar in spirit to the spatial expansion model of Casetti (1972), in which a linear model's coefficients are expressed as function of spatial data set's geographic coordinates. The model is considered in detail by Cleveland, Grosse, and Shyu (1992) and Cleveland (1994). A more general version of the model is also analyzed in Hastie and Hibshirani (1992).

Although the CPAR model is similar to the semiparametric model (Robinson, 1988) in that some explanatory variables are fully nonparametric while others have parametric, the CPAR model is more general than the semiparametric model. The semiparametric model is

$$y_i = \beta_1(X_1) + \beta_2' X_2 + u \tag{6}$$

which differs from the CPAR model in that the β_2 is constrained not to vary with X_2. The approaches can be combined by allowing one set of variables to be fully parametric, another to be fully nonparametric, and a third to be conditionally parametric on the second set of variables.

The objective function for the CPAR model is a modification of equation (1):

$$\sum_{i=1}^{n}\left(y_i - \alpha - \beta_1'(X_{1i} - X_1) - \beta_2' X_{2i}\right)^2 K\left(\frac{X_{1i} - X_1}{h}\right) \tag{7}$$

Equation (6) leads to the following estimates:

$$\hat{\theta}(X) = \left(\sum_{i=1}^{n} K(\psi_{1i}) Z_i Z_i' \right)^{-1} \sum_{i=1}^{n} K(\psi_{1i}) Z_i' y_i \qquad (8)$$

where $\psi_{1i} = \left(\frac{X_{1i} - X_1}{h} \right)$; Z includes a constant, X_1, and X_2; and $\theta = (\alpha \quad \beta_1 \quad \beta_2)'$.

Thus, the model is exactly the same as LWR except that the kernel weight function only includes the variables in X_1. Distance is defined in terms of X_1 alone, but all variables are included in the WLS regression.[5]

The CPAR is a natural candidate for many spatial data sets. For example, in a hedonic study we might expect house prices to be linear in the structural characteristics at any given location, but the intercept and marginal effects may vary over space. In this case, a natural specification is to define X_1 to include latitude and longitude while X_2 includes the structural characteristics. Although locational variables such as distance from the city center are implicit functions of X_2 in this situation, they can also be included in X_1 if their marginal effects are conditionally parametric at any given location. In this case, the kernel function might be specified as $K\left(\frac{z_{1i} - z_1}{h} \right) K\left(\frac{z_{2i} - z_2}{h} \right)$, where z_1 and z_2 are the geographic coordinates standardized to have unit variance. This kernel function is used to define the weights for a WLS regression of y on X_1, z_1, and z_2. The estimated coefficients for all variables with vary over space, but only the geographic coordinates are used to define target points for estimation.

---

[5] An alternative way to derive the same estimator is discussed in McMillen (2004a). When estimating multivariate models, most researchers use product kernels, such as $K(\psi) = K_1(\psi_1) K_2(\psi_2)$ in the two-variable case. If a rectangular kernel is used for X_2 with the window width set to the maximum of all available observations, then the LWR estimator simplifies to the CPAR model. This specification of the kernel is appropriate if the function is nonlinear in X_1 while being conditionally linear in X_2, which is the assumption behind the CPAR model.

What has come to be known as geographically weighted regression (GWR) is a special case of this special case of LWR estimation. Rather than using a general function of latitude and longitude to define the kernel weights, the GWR weights are based on the straight-line distance between observation *i* and the target point (d_i) . In addition, the geographic coordinates are typically omitted from the list of explanatory variables. Thus, the objective function for GWR estimation is the following modification of equation (7);

$$\sum_{i=1}^{n}\left(y_i-\alpha-\beta_2{}'X_{2i}\right)^2 K\left(\frac{d_i}{h}\right) \tag{9}$$

The model is estimated using a WLS regression of *y* on a constant and the variables in the vector X_2. The GWR model appears to have first been used in McMillen (1996) and Brunsdon, Fotheringham, and Charlton (1996). It has since been used in a series of papers by both sets of authors, although McMillen uses the term LWR instead of GWR to recognize that it is merely an application of the procedure developed originally by Cleveland and Devlin (1988). It is limiting to view GWR as an estimator in its own right; it is a special case of LWR and CPAR estimation that only allows spatial variation in marginal effect estimates. It does not allow for nonlinearities in other variables, and it uses a special kernel that is based on a circle around the target point, which limits its application to urban housing submarkets which can be organized in decidedly asymmetric shapes. Recognizing its status as a special case makes it easier to consider useful generalizations while providing a link to other literatures.

3. Hypothesis Testing in Nonparametric Estimation

As target points for nonparametric estimation typically include every data point, a grid covering the relevant geographic area, or simply a set of interesting points, LWR hypothesis testing is more nuanced than is the case with conventional estimators.[6] Whereas standard hedonic approaches produce one population parameter, the outcome of the LWR is a set of estimates *for each regressor.* With conventional estimators, standard hypothesis testing involves asking whether the regressor adds significantly to the explanatory power of the model – whether *the* coefficient associated with it is significantly different than zero. With nonparametric models – because each target point has a different set of coefficient estimates – two types of hypothesis testing are possible. The first is the analogue to traditional hypothesis testing, regarding the added explanatory power of a regressor. The second is more localized, asking *where* the regressor is statistically significant.

The analog to traditional hypothesis testing is straightforward. Covariance matrix estimates are available for each estimate, but a full covariance matrix for all estimated coefficients is hard to construct because the estimates are not independent across target points. Nonetheless, Cleveland and Devlin (1988) show that good approximations are available for tests of the null hypothesis that a variable (or set of variables) adds no explanatory power to a nonparametric. The test is comparable to a standard F-test, and has nearly the same form.

[6] Although nonparametric models are often estimated with each observation as a target point, estimation time can be reduced substantially by estimating the model over a grid of target points or for a subset of the observations. Loader (1999) discusses alternative interpolation procedures for constructing estimates at each observation when the number of target points is less than the number of observations.

The key to constructing Cleveland and Devlin's *F*-test is to note that all of the estimators discussed in the previous section – OLS, LWR, kernel regression, CPAR, and GWR – have the same linear form $Y = LY + u$, where Y is the nx1 vector of explanatory variable values, u is the vector of residual terms, and L is an nxn matrix. The vector of residuals is $u = (I - L)Y$ and the residual sum of squares is $u'u = Y'(I - L)'(I - L)Y$, which can also be written as $Y'RY$, where $R = (I - L)'(I - L)$. In the standard linear regression model, $L = X(X'X)^{-1}X'$ and $(I - L)$ is a symmetric, idempotent matrix such that $(I - L)'(I - L) = (I - L)$. This feature of the OLS estimator greatly facilitates hypothesis testing because it implies that the variance will follow a central χ^2 distribution when the errors are normally distributed.

The statistical theory is more complicated for the nonparametric estimators because $(I - L)$ is neither symmetric nor idempotent. However, Cleveland and Devlin (1988) note that the distribution can be approximated readily. Let R_r represent the value of $(I - L)'(I - L)$ under the null (or "restrictive") model (e.g., with one variable deleted from the list of explanatory variables), and let R_a the value under the alternative. Next, define the following values: $\delta_1 = tr(R_a)$, $\delta_2 = tr(R_a R_a)$, $\upsilon_1 = tr(R_r - R_a)$, and $\upsilon_2 = tr\left[(R_r - R_a)(R_r - R_a)\right]$. The counterpart to the standard F-test is then:

$$\frac{(Y'R_rY - Y'R_aY)/\upsilon_1}{(Y'R_aY)/\upsilon_2} \sim F\left(\upsilon_1^2/\upsilon_2, \delta_1^2/\delta_2\right) \qquad (10)$$

Examples of the application of this approach include McMillen (1996) and McMillen and McDonald (1997). In small samples, the degrees of freedom parameters can be

calculated directly, but δ_2 and (particularly) υ_2 are difficult to calculate for large samples because they require the multiplication of large matrices.

Fortunately, Loader (1999) shows that a simple approximation is quite accurate for the degrees of freedom calculation. Define $d_1 = tr(L)$ and $d_2 = tr(L'L)$. Then the degrees of freedom used in estimation by the nonparametric estimator is approximately equal to $\kappa = 2d_1 - d_2$. Using this simplification, the F-test becomes simply:

$$\frac{\left(Y'R_rY - Y'R_aY\right)/\left(\kappa_a - \kappa_r\right)}{\left(Y'R_aY\right)/\left(n - \kappa_a\right)} \sim F\left(\kappa_a - \kappa_r, n - \kappa_a\right) \tag{11}$$

where κ_r and κ_a are the degrees of freedom used in the null and alternative models. This F-test has nearly the identical form as a standard F-test, and all of the terms are easy to calculate. To form counterparts to standard t-statistics, equation (11) can be calculated with each variable dropped from the model. The prob-values from these tests are analogous to the results of a standard t-test of the null hypothesis that the coefficients equals zero. In other words, equation (11) can be used to test whether an explanatory variable (or a set of variables) adds any explanatory power to the nonparametric model.

Just as hypothesis tests of linear regression models are conditional on the model specification, the results of a nonparametric model are conditional on the list of explanatory variables and the bandwidth or window size. The voluminous literature on bandwidth selection is still growing. A common method for choosing the bandwidth or window size is *cross validation*: each observation is used as the target point, and the estimated value of *y* for observation *i* is constructed after omitting the *i*th observation from the model. The computer-intensive cross-validation approach is similar to out of sample forecasting. Following work by Craven and Wahba (1979), Loader (1999) shows

that the less computer-intensive generalized cross-validation score (GCV) closely approximates the results of the cross-validation approach:

$$GCV = n\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\left(n - d_1\right)^2} \tag{12}$$

where $\hat{y}_i$ is the predicted value of y_i and $d_1 = \text{tr}(L)$. The degrees of freedom correction penalizes smaller bandwidths without requiring the model to be re-estimated for each observation.

Equation (11) does not produce confidence intervals for a variable's marginal effect on the dependent variable. Simple summary statistics such as means and quintiles or simply maps can be used to summarize the coefficient estimates from the nonparametric models. Alternatively, bootstrap procedures can be used to construct confidence intervals for any statistics calculated after estimating the model. Bootstrap standard errors estimates are particularly easy to calculate with procedures such as the wild bootstrap of Härdle (1990) because once the matrix L is calculated, new estimates can be constructed trivially with repeated bootstrap samples of y. McMillen (2004b) uses this estimator in a LWR analysis of employment densities.

In this paper, we use semi-parametric estimation as an alternative method of summarizing the average effect of a single variable. When a single variable z enters the model parametrically, the estimating equation can be written as $y_i = f(X_i) + \gamma z_i + u_i$. Following Robinson (1988), estimation proceeds in the following steps: (1) use any of the nonparametric estimators to regress y on X and each z on X, form the residuals (e_y and e_z), and (2) use OLS to regress e_y on e_z. The coefficient on e_z in the second-stage linear regression is the estimate of γ, and the standard error from the regression can be

used to form confidence intervals. The semi-parametric model provides an estimate of the conditional expectation of y given z after controlling in a general, nonparametric way for the effects of all other variables. Semi-parametric estimation is a simple combination of nonparametric and linear regression procedures, and it provides a direct analog to the standard parametric coefficient and standard error of a linear model.

4. Monte Carlo Results

A series of limited Monte Carlo experiments illustrates some of the features and advantages of LWR estimation. The basis for the experiments is a stylized model of an urban area in which the dependent variable declines with distance from the city center. All observations are arranged along a single line running through the city center. Points along the line (n = 2000) are drawn from a uniform distribution ranging from x = -20 to x = 20, i.e., a U(-20,20) distribution. The base function is linear for $x < 0$, with $y = 11.25 + .5x + u$. The function is more complicated for $x>0$: $y = 10 + 1.25\sin(z) + 1.25\cos(z) - .5x + .5x^2/1000 + u$, where $z = 2\pi x / 20$. The base function is shown as the solid line in Figure 1. It looks much like a population density function in a monocentric city in which log-densities decline much more rapidly with distance from the city center on the east side of the city. All estimates are based on a single draw of errors from a normal distribution with a mean of zero and a variance of 16.67. The value for the variance implies that the R^2 will be approximately 0.67 when the function is correctly specified.

The dotted line in Figure 1 shows the estimated line from a LWR model with y as the dependent variable, $|x|$ (i.e., distance from the city center) as the explanatory variable, and weights based on the geographic distance between each observation j and the target

point i, or $\left|x_i - x_j\right|$. Thus, this model can be thought of as a LWR, CPAR, or GWR model. We use a tricube weight function with the window size chosen by minimizing the value of GCV. The GCV criterion indicates that 20% of the observations should be given positive weight in estimating the value of y at any target point x_i. The estimates closely approximate the true line even though the base model is linear in x. The functions are somewhat noisy west of the city center where the true function is linear. This problem could be eliminated by choosing a larger size on the west side.

Figure 2 shows that the form of the kernel weight function has virtually no effect on the results. Six different kernels are shown – the rectangular, triangular, Epanechnikov, bisquare, tricube, and triweight kernel. In each case, the window size is chosen by minimizing the GCV value. The estimated values are so close across kernels that it is virtually impossible to discern one from another.

Figure 3 shows that the estimates are affected by the window size. When the window size is set to 50% rather than the 20% indicated by the GCV, the estimated function is close to being linear on both sides of the city. The estimates approximate the true values well on the east side, but they lead to too much smoothing on the west side. The estimates are noisy on both sides of the city when the window size is 10%.

Optimal window sizes are larger when the goal is to estimate marginal effects rather than to simply predict the dependent variable accurately (Pagan and Ullah, 1999). Figure 4 illustrates this point using the LWR estimator evaluated at two window sizes, 20% and 40%. The solid line shows the marginal effect of distance from the city center based on the true underlying function. When the function is close to linear (for $x<0$), the 20% window size produces noisy derivative estimates even though the GCV criterion

indicated that 20% is the optimal window size for predicting y. The smaller window size appears less noisy on the east side of the city where the function is nonlinear. Over the full function, it appears that a window size of 40% is more accurate than the smaller window.

It is worth emphasizing that the only information provided to the LWR estimator is that y is a function of x. With no knowledge other than the correct explanatory variable, the LWR model does an impressive job of tracking a fairly complicated function. The GCV version of cross validation is easy to calculate and provides a useful guide for choosing the window size when the goal is to predict the dependent variable accurately. The optimal window size is much larger – perhaps double – when the goal of estimation is to estimate the marginal effect of the explanatory variables.

The base parametric model implicitly uses six degrees of freedom – a constant, x, and interaction terms between an east side dummy variable and x, x^2, $\sin(z)$, and $\cos(z)$. (It would use seven degrees of freedom if the constant term were not constrained to be equal across the two sides of the city.) The first column of results in Table 1 shows the number of degrees of freedom that are used to estimate the CPAR version of the model in which $|x|$ is the single explanatory variable and x is used to form the tricube kernel weights. With the GCV window size of 20%, the LWR model uses approximately 11 degrees of freedom to obtain virtually the same level of predictive accuracy as the underlying parametric function. Even a very narrow 10% window size uses only about 21 degrees freedom. For this simple model with one explanatory variable, the degrees of freedom are about the same for both the CPAR and kernel regression models.

To compare the LWR model with a fixed effects specification, we need to form the Monte Carlo counterparts to geographic areas such as census tracts. Since census tracts comprise areas with approximately equal populations, a natural choice for the Monte Carlo experiments is a series of equally space intervals along the *x*-axis of the diagrams. In practice, the number of geographic areas would likely be determined by political jurisdictions or census boundaries. Here we have the luxury of allowing the data to choose the optimal number of fixed effects, which places the fixed effects estimator in its best possible light. Using the GCV criterion to choose the number of intervals for the *x*, the optimal fixed effects estimator uses 15 degrees of freedom to estimate the model – an intercept, distance from the CBD, and 13 dummy variables indicating the interval for *x*. Thus, even an optimal number of fixed effects uses more degrees of freedom than the LWR model. Figure 5 shows why: the fixed effects estimator is forced to approximate a smooth function with a series of line segments with identical slopes. While a piecewise linear function can approximate a smooth surface, it is less accurate than a flexible functional form that does not impose an incorrect specification with marked discontinuities.

The remaining columns of Table 1 show how the degrees of freedom vary as the number of explanatory variables increases. The additional explanatory variables are drawn independently from U(0,1) distributions. The base model remains the same as before since the point is simply to illustrate how degrees of freedom increase as the number of explanatory variables increases in estimation.[7] The number of degrees of

[7] For the CPAR model, the tricube function remains a function of *x* alone, but the new variables are included as explanatory variables. For the kernel regression model, both *x* and the additional variables are included as explanatory variables and in the tricube weight function.

freedom increases rapidly as the number of explanatory variables increases and the window size declines. CPAR greatly conserves on degrees of freedom by confining the additional variables to the conditionally parametric part of the model. These results show that (1) nonparametric estimators are not profligate users of degrees of freedom when window sizes are in the standard 20%-50% range, and (2) the CPAR method has significant advantages for spatial modeling when the primary source of nonlinearity is spatial variation in the coefficients of an underlying linear model.

5. Hedonic Estimates

In this section, we illustrate the benefits of nonparametric estimation in a typical hedonic price function setting. The data set and empirical question are similar to those in the study by McMillen and McDonald (2004): how do house prices vary with proximity to a rapid transit line? The data set includes all sales of single family homes in 2000 that were within one mile of Chicago's elevated train line (the "EL"). The sales data are merged with assessor's data to obtain standard housing characteristics, including building area; lot size; the number of rooms, bedrooms, and bathrooms; and dummy variables indicating that the home has a brick exterior, fireplace, central air conditioning, and a garage. After geo-coding the data, we used a GIS program to measure distance from Chicago's city center (at the intersection of State and Madison streets) and distance from the nearest EL stop. We also constructed dummy variables indicating that a home is within a block of a rail line or an EL line. While access to a stop on the EL is presumably an amenity, the noise associated with a location close to the line itself is likely to lower

prices. Rail and EL lines often run through commercial and manufacturing areas, which also may act as disamenities for homeowners.

Descriptive statistics are shown in Table 2. The base hedonic specification is Model 1 in Table 3. With the natural logarithm of sales price as the dependent variable and a combination of structural and locational characteristics as explanatory variables, this specification is entirely standard with the exception of longitudes and latitudes. Including the geographic coordinates accounts for broad geographic trends and serves as a link to the general version of the LWR model. The base model explains a respectable 63.1% of the variation in log-sales prices. Most of the results are standard, with prices increasing with square footage, lot size, and with the presence of such amenities as fireplaces, central air conditioning, and a garage. The important result for our purpose is the coefficient on distance from the nearest EL stop: prices are approximately 13.7% lower one mile from an EL stop than at the station. This is the figure that is used to construct a measure of the potential benefit of building a new EL line or opening a new stop (Cervero 1997, 2001). Hedonic estimates can be sensitive to the specification, particularly to the specification of locational effects. A common method of controlling for local fixed effects is the inclusion of census tract dummy variables. Model 2 of Table 3 adds 466 census tract fixed effects to the base regression. Not surprisingly, the results change substantially. Although the R^2 rises to 0.870, several variables that had been quite significant drop to insignificance. After controlling for census tract fixed effects, the point estimate for distance to the nearest EL stop falls to -0.044 and it is not statistically different from zero. However, 483 degrees for freedom are required for this model, or about 13% of the number of observations in the data set. Several of the census

tracts have only one or two observations. In a Monte Carlo setting, the estimator represented by Model 2 would have a very high variance.

There is an additional point to be made here about common practice and the use of Census tract fixed effects. It is not uncommon to see Census fixed effects used to absorb omitted spatial variables. The problem with this approach is that Census tracts are fixed in space – broad spatial patterns will be lost in regressions that use that use these fixed effects. In the fixed effects regression above, not only are 483 degrees of freedom used, the distance variable now captures *within-tract* variation in housing price due to proximity. Given the irregular shapes of Census tracts and potential for broad non-linearities across them, it is no surprise that the distance variable is not significant. However, it would be inappropriate to conclude that distance is not a significant from such regressions; rather, an auxiliary regression would be needed that regressed Census tract fixed effects on the distance from the tracts to the stations. This step is generally not undertaken.

Tables 4 and 5 show CPAR and semi-parametric estimates at two window sizes, 25% and 100%. All of the models are estimated using a tricube kernel function with straight-line distance between observations as its sole argument. Each observation is used as a target point for estimation. This approach is similar in spirit but more general than the fixed effects specification. While it allows for spatial variation in the house price surface by including the geographic coordinates as explanatory variables, it also allows all estimated coefficients to vary smoothly over space. Separate semi-parametric models are estimated for each explanatory variable. The listed explanatory variable is constrained to enter the model parametrically, while all other variables enter the model in

CPAR form with spatially varying coefficients. The tables show (1) the average coefficients from the nonparametric CPAR models and (2) parametric coefficients from the semi-parametric models. Sample standard deviations are listed for the CPAR models, along with the Cleveland-Devlin (1988) F-test values for the null hypothesis that the variable adds no explanatory power to the regression.

Both Tables 4 and 5 suggest that the coefficients for the structural characteristics are quite stable on average and produce results that are similar to their counterparts in Table 4. For example, the average coefficient for the log of building area is 0.315 across the 3705 CPAR estimates when the window size is 25%, with a standard deviation of 0.088. The semi-parametric estimate is 0.323 with a standard error of 0.026. The difference between the two sets of estimates is that all coefficients vary over space with the CPAR model, including the coefficient for the log of building area. The semi-parametric model produces only a single estimate for the coefficient for the log of building area, but still yields 3705 estimates of the coefficients for all other variables. Comparable values for these estimates are 0.383 and 0.379 when the window size is increased to 100% in Table 5. The estimates can be compared to the linear regression models of Table 3, in which the base estimated coefficient of 0.403 falls to 0.281 when the model includes census tract fixed effects. The 25% window size produces results comparable to the fixed effects specification, while the 100% window size produces results similar to the linear model without fixed effects. For the log of building area, the primary result of geographic disaggregation is to produce a lower value for the estimated marginal effect.

The results are more sensitive to the specification for the locational variables – longitude, latitude, distance from the city center, distance from an EL Stop, and the dummy variables indicating that the observation is within a block of a rail line or an EL line. This result should not be surprising. By definition, the locational variables are themselves functions of the very geographic coordinates that determine distances between observations. As the window size collapses toward zero, the locational variables come closer to being constants within the window of observations received positive weight at each target point. This degeneracy leads to nearly perfect collinearity at small window sizes. The same problem helps explain the sensitivity of these variables to the presence of census tract dummy variables in the fixed effects specification. Broad spatial effects such as distance from the city center or distance from the nearest EL stop require fairly large window sizes in order to separate their effects from the effects of other spatial variables.

The Monte Carlo results suggested that the optimal window size is larger when the goal is to estimate marginal effects than when the objective is prediction. In addition, the fact that small window sizes (or small neighborhoods) lead to highly variable results for locational variables suggests that still larger window sizes should be used when the goal is to estimate the marginal effects of locational variables than when the goal is to estimate the marginal effects of structural characteristics. Thus, the 100% window size is likely preferable to the 25% window for estimating the marginal effect of a broad geographic variable such as distance to the city center or distance to the nearest El stop.

Based on the CPAR model with a 100% window size, the results for our primary variable of interest suggest that prices decline on average by nearly 17% per mile with

distance from the nearest EL stop. The standard deviation for the mean coefficient for distance from the nearest EL stop is 0.028 and the Cleveland-Devlin (1988) F-test implies that the variable adds significant explanatory power to the regression. The semi-parametric result suggests that the magnitude of the marginal effect of this variable is lower at -10.9%, but still highly significant. The reason for the difference between the two sets of estimates can be seen in Figure 6, which shows the estimated CPAR coefficients for the smaller, 25% window size.[8] The coefficients are clearly negative for most observations north of the city center, meaning that prices are higher closer to the EL stops. In accordance with the results of McMillen and McDonald (2004), the estimated coefficients are also negative along the line running southwest from the city center. Positive coefficients are confined to two areas – due south of the city center and on the west side. Figure 7 presents a histogram showing the distribution of coefficients across all observations. Though the number of negative coefficients is much higher than the number of positive coefficients, it is clear that the EL line is not an amenity everywhere.

The south and west sides of the Chicago are relatively low-income areas, and the areas near the EL stops tend to be blighted commercial areas with many vacant buildings. The map shows that combining all EL lines into the same sample produces a serious form of model misspecification. This misspecification would not have been evident in the base linear specification. Thus, an important advantage of nonparametric modeling is the information it provides for model testing. In the case of Chicago's EL, the coefficients on access to rail could be used to examine the influence of income, race/ethnicity, family size, labor force attachment, human capital, and other fundamental variables on the

[8] The geographic variation in the coefficients shows up more clearly on the map when the window size is 25%, but the 100% window-size results are similar.

demand for mass transit – research opportunities not generally possible using standard hedonic analysis.

When comparing the CPAR and linear results, note that nonparametric estimation does not use up an inordinate number of degrees of freedom even though the models are constructed by estimating separate WLS regressions for each observation. With a 100% window size, only about 28 degrees of freedom are used to estimate a model that starts out with 18 explanatory variables. Ten additional degrees of freedom produce an impressive degree of spatial variation in the estimated coefficients. Even with the much smaller 25% window, the CPAR model uses far fewer degrees of freedom than the census tracts fixed effects models – 164 v. 483. The commonly held impression that all nonparametric estimators lead to very high variances in models with more than one or two explanatory variables is simply wrong. Many degrees of freedom are saved by taking advantage of the special nature of spatial data sets to include only distance in the kernel weight function and imposing – mechanically – smoothness in the surface of parameter estimates.

The second point to note when comparing CPAR and linear results is that the hypothesis tests provide virtually the same information. The coefficient of a linear model can be interpreted as an estimate of the average marginal effect of an explanatory variable. We have accomplished the same objective for nonparametric models by calculating sample averages for the CPAR model and by estimating semi-parametric versions of the models. Cleveland and Devlin's (1988) F-test is an effective way of summarizing a variables overall statistical significance. The standard error for a semi-

parametric estimate provides exactly the same information as its counterpart from a linear regression.

Finally, we have shown the economic relevance of the hedonic price surfaces. Rather than being an obstacle to interpretation, they are useful data on the spatial distribution of fundamentals and preferences that produce outcomes in housing markets. We have illustrated the spatial variation in the pricing of the access to rail service and suggested that there is a economic rationale for the asymmetric pricing that may serve as a far better input into the policy process that is asked to evaluate the social costs and benefits of public goods than a single population parameter.

6. Conclusion

Nonparametric estimation procedures can control for spatial variation in marginal effects while also allowing for nonlinearities. Variants of the LWR procedure are easy to implement because they simply require repeated applications of standard WLS regressions. Although the typical procedure involves a separate regression for each point in the data set, LWR procedures often use fewer degrees of freedom than regressions with neighborhood fixed effects. The WLS regressions are simply a convenient way to construct the nonparametric estimates; the restrictions imposing continuity over neighboring observations are sufficient to conserve degrees of freedom. Still fewer degrees of freedom are used when the model is conditionally parametric. In the conditionally parametric version of LWR, the base model is linear conditional on the variables used to construct the WLS weights, but the coefficients can vary smoothly across nearby observations. In spatial data sets, “nearby” has a natural geographic

interpretation, and the conditionally parametric model allows marginal effects to vary smoothly over space. This spatial interpretation of the LWR estimator is sometimes referred to as "geographically weighted least squares," although the term is unfortunate because it conceals the estimator's relationship to more general procedures that can also be very useful in a spatial context.

After illustrating the relationships among various LWR estimators analytically and in a set of Monte Carlo experiments, we demonstrate the usefulness of LWR estimation in a representative hedonic setting. Using data on sales of single-family homes in Chicago for 2000, we find that the estimates of the marginal effects of such spatial variables as distance from the nearest EL stop are sensitive to the model specification. Prices are estimated to decline significantly with distance to an EL stop when the parametric model only includes such broad measures of spatial trends as longitude, latitude, and distance from the city center. Parametric neighborhood fixed effects or nonparametric LWR estimates reduce the level of significance substantially, and the marginal effect of distance to the nearest EL stop can be reduced to statistical insignificance if neighborhoods are defined narrowly. This sensitivity of location variables to the neighborhood definition is shared by both parametric and LWR models.

One of the advantages of LWR estimation comes at the next stage of determining *why* the estimates are sensitive to the neighborhood definition. A map of the LWR coefficients for distance to the nearest EL stop reveals that proximity to the EL significantly increases sales prices on the north and southwest sides of the city. In low-income areas on the south and west sides, the EL is lined with vacant businesses and other depreciated structures. Failure to account for these differences leads to an

underestimate of the value of proximity to the EL in higher-income areas. Our empirical application also illustrates that hypothesis testing is not much more difficult in a nonparametric setting than in a conventional regression model. Accurate approximations are available for F-tests that provide virtually the same information as a standard t-statistic, and semiparametric models can be used to construct confidence intervals for the marginal effect of individual explanatory variables.

References

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

C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically Weighted Regression," *Geographical Analysis*, 28 (1996) 281-298.

E. Casetti, "Generating Models by the Expansion Method: Applications to Geographical Research," *Geographical Analysis*, 4 (1972) 81-91.

R. Cervero and M. Duncan, "Transit's Value-Added: Effects of Light and Commuter Rail Services on Commercial Land Values," University of California's Institute of Transportation Studies Working Paper (2001).

R. Cervero and K.-L. Wu , "Polycentrism, Commuting, and Residential Location in the San Francisco Bay Area," *Environment and Planning A*, 29 (1997) 865-886.

W. S. Cleveland, "Coplots, Nonparametric Regression, and Conditionally Parametric Fits," in T. W. Anderson, K. T. Fang, and I. Olkin (Eds.), *Multivariate Analysis and its Applications* (Hayward: Institute of Mathematical Statistics, 1994).

W. S. Cleveland and S. J. Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83 (1988) 596-610.

W. S. Cleveland, E. H. Grosse, and W. M. Shyu, "Local Regression Models," in J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S* (Pacific Grove: Wadsworth and Brooks/Cole, 1992).

P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik*, 31 (1979) 377-403.

A. C. Goodman and T. G. Thibodeau, "The Spatial Proximity of Metropolitan Area Housing Submarkets," *Real Estate Economics***,** 35 (2007) 209-232

A. C. Goodman and T.G. Thibodeau, "Housing Market Segmentation," *Journal of Housing Economics*, 7 (1998) 121-143.

R. Halvorsen and H. O. Pollakowski, "Choice of Functional Form for Hedonic Price Equations," *Journal of Urban Economics*, 10 (1981) 37-49.

W. Härdle, *Applied Nonparametric Regression* (New York: Cambridge University Press, 1990).

T. J. Hastie and R. J. Tibshirani, "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Series B,* 55 (1992) 757-796.

C. Loader, *Local Regression and Likelihood* (New York: Springer, 1999).

D. P. McMillen, "One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach," *Journal of Urban Economics*, 40 (1996) 100-124.

D.P. McMillen, "Employment Subcenters and Home Price Appreciation Rates in Metropolitan Chicago," in J. P. LeSage and K. Pace (Eds.), *Advances in Econometrics, Volume 18: Spatial and Spatiotemporal Econometrics* (New York: Elsevier, 2004a).

D. P. McMillen, "Employment Densities, Spatial Autocorrelation, and Subcenters in Large Metropolitan Areas," *Journal of Regional Science*, 44 (2004b) 225-243.

D. P. McMillen and J. F. McDonald, "A Nonparametric Analysis of Employment Density in a Polycentric City," *Journal of Regional Science,* 37 (1997) 591-612.

D. P. McMillen and J. F. McDonald, "Reaction of House Prices to a New Rapid Transit Line: Chicago's Midway Line," *Real Estate Economics*, 32 (2004) 463-486.

R. Meese and N. Wallace, "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 (1991) 308-332.

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

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

S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, 82 (1974) 34-55.

J. H. Stock, "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in W. A. Bennett, J. Powell, and G. E. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (New York: Cambridge University Press, 1991).

Table 1

Degrees of Freedom

Window Size	Number of Explanatory Variables					
	1	2	3	4	5	6
Conditionally Parametric Regression						
0.10	21.506	41.631	61.632	81.647	101.547	121.468
0.20	11.298	21.236	31.125	41.023	50.948	60.854
0.30	7.895	14.424	20.950	27.452	33.984	40.532
0.40	6.194	11.011	15.849	20.667	25.498	30.355
0.50	5.173	8.968	12.781	16.588	20.411	24.244
0.60	4.494	7.610	10.744	13.872	17.018	20.161
0.70	4.034	6.663	9.310	11.953	14.615	17.266
0.80	3.721	5.989	8.270	10.552	12.855	15.141
0.90	3.500	5.487	7.484	9.487	11.509	13.510
1.00	3.310	5.072	6.844	8.620	10.416	12.191
Kernel Regression						
0.10	21.158	47.353	94.265	177.495	302.917	481.661
0.20	10.879	23.272	40.130	62.388	90.875	126.269
0.30	7.637	15.795	25.049	35.348	47.412	61.060
0.40	5.763	11.603	17.800	24.508	31.759	39.509
0.50	4.796	9.400	14.147	19.133	24.379	29.856
0.60	4.230	8.027	11.847	15.791	19.865	24.050
0.70	3.581	6.632	9.679	12.775	15.930	19.151
0.80	3.208	5.818	8.421	11.048	13.704	16.408
0.90	2.948	5.236	7.519	9.813	12.121	14.470
1.00	2.753	4.794	6.825	8.863	10.911	12.987

Table 2
Descriptive Statistics

	Mean	Std. Dev.	Minimum	Maximum
Sales Price	223250	169333	20000	1015000
Building area (s.f.)	1365	608	402	10041
Land area (s.f.)	3533	1358	478	18326
Log of sales price	12.070	0.710	9.903	13.830
Log of building area	7.142	0.376	5.996	9.214
Log of land area	8.087	0.441	6.170	9.816
Age of Structure	69.979	35.091	1	138
Rooms	5.744	1.593	3	15
Bedrooms	2.923	0.851	1	8
Bathrooms	1.506	0.704	1	8
Brick exterior	0.552	0.497	0	1
Fireplace	0.170	0.375	0	1
Central air conditioning	0.283	0.451	0	1
Garage	0.307	0.461	0	1
Within a block of a rail line	0.256	0.436	0	1
Distance from city center (miles)	6.960	2.613	0.849	12.699
With a block of an EL line	0.099	0.299	0	1
Distance from EL stop	0.553	0.251	0.015	1.000
Longitude	-87.69808	0.05087	-87.83994	-87.58022
Latitude	41.87923	0.08589	41.69902	42.01892

Note. The data set comprise 3705 sales of single-family residential homes in Chicago for 2000.

Table 3
Linear Regression Results

	Model 1			Model 2		
	Coef.	Std. Err.	P-value	Coef.	Std. Err.	P-value
Log of building area	0.403	0.034	0.000	0.281	0.025	0.000
Log of land area	0.186	0.024	0.000	0.264	0.019	0.000
Age of Structure	-0.002	0.0003	0.000	-0.001	0.0002	0.020
Rooms	0.021	0.009	0.015	-0.001	0.006	0.915
Bedrooms	-0.024	0.015	0.097	0.015	0.010	0.133
Bathrooms	0.030	0.017	0.073	-0.008	0.012	0.491
Brick exterior	0.093	0.016	0.000	0.026	0.012	0.036
Fireplace	0.179	0.024	0.000	0.022	0.018	0.218
Central air conditioning	0.136	0.021	0.000	0.032	0.015	0.027
Garage	0.027	0.016	0.082	0.025	0.011	0.021
Within a block of a rail line	-0.007	0.018	0.708	0.030	0.015	0.054
Distance from city center	-0.032	0.004	0.000	0.028	0.040	0.481
With a block of an EL line	0.006	0.029	0.831	-0.019	0.025	0.453
Distance from EL stop	-0.137	0.033	0.000	-0.044	0.041	0.291
Longitude	1.644	0.213	0.000	1.978	2.242	0.378
Latitude	4.512	0.118	0.000	1.340	2.295	0.559
Constant	-36.864	15.863	0.020	125.096	218.592	0.567
R^2	0.631			0.870		
Degrees of freedom	17			483		

Notes. Model 1 is the base specification. Model 2 adds 466 census tract fixed effects to the model. The F-statistic for the fixed effects is F(466, 3222) = 12.629, with a p-value of 0.000.

Table 4

CPAR Results:
Window Size = 25%

	Nonparametric				Semi-parametric		
	Mean Non-Param.	Std. Dev.	F-Test	P-Value	Semi-Param.	Std. Err.	P-Value
Log of building area	0.315	0.088	16.527	0.000	0.323	0.026	0.000
Log of land area	0.216	0.096	13.851	0.000	0.219	0.020	0.000
Age of Structure	-0.002	0.002	13.893	0.000	-0.001	0.0002	0.000
Rooms	0.004	0.019	1.441	0.149	0.001	0.007	0.923
Bedrooms	-0.006	0.028	1.548	0.108	-0.004	0.011	0.734
Bathrooms	0.030	0.066	3.400	0.000	0.020	0.013	0.115
Brick exterior	0.035	0.055	3.723	0.000	0.039	0.013	0.002
Fireplace	0.065	0.078	4.567	0.000	0.081	0.019	0.000
Central air conditioning	0.054	0.061	2.811	0.001	0.047	0.016	0.003
Garage	0.031	0.037	2.051	0.020	0.030	0.012	0.012
Within a block of a rail line	-0.019	0.049	2.437	0.007	-0.017	0.014	0.226
Distance from city center	-0.346	0.522	31.118	0.000	-0.058	0.030	0.050
With a block of an EL line	0.001	0.076	1.052	0.396	-0.028	0.023	0.236
Distance from EL stop	-0.049	0.147	4.533	0.000	-0.016	0.026	0.545
Longitude	-7.678	15.171	25.380	0.000	5.010	0.952	0.000
Latitude	17.069	31.251	45.034	0.000	9.758	1.124	0.000
R^2	0.825						
Degrees of freedom	164.144				141.024		

Notes. The semi-parametric estimates are the values from individually estimated models. The degrees for freedom for the semi-parametric models are the average across the estimated models.

Table 5

CPAR Results:
Window Size = 100%

	Nonparametric				Semi-parametric		
	Mean Non-Param.	Std. Dev.	F-Test	P-Value	Semi-Param.	Std. Err.	P-Value
Log of building area	0.383	0.042	82.489	0.000	0.379	0.033	0.000
Log of land area	0.175	0.015	35.027	0.000	0.176	0.023	0.000
Age of Structure	-0.002	0.001	78.685	0.000	-0.002	0.0003	0.000
Rooms	0.028	0.012	6.705	0.002	0.017	0.008	0.046
Bedrooms	-0.029	0.008	2.094	0.128	-0.024	0.014	0.090
Bathrooms	0.019	0.019	8.242	0.001	0.040	0.016	0.014
Brick exterior	0.087	0.022	24.617	0.000	0.088	0.015	0.000
Fireplace	0.179	0.004	31.570	0.000	0.172	0.023	0.000
Central air conditioning	0.152	0.010	17.432	0.000	0.121	0.020	0.000
Garage	0.028	0.002	3.024	0.053	0.039	0.015	0.010
Within a block of a rail line	-0.013	0.005	0.708	0.477	-0.014	0.017	0.424
Distance from city center	-0.029	0.017	74.598	0.000	-0.018	0.004	0.000
With a block of an EL line	0.009	0.012	0.026	0.956	0.011	0.028	0.699
Distance from EL stop	-0.169	0.028	7.056	0.001	-0.109	0.032	0.001
Longitude	1.832	0.468	72.654	0.000	2.181	0.275	0.000
Latitude	4.641	0.199	1455.817	0.000	5.545	0.156	0.000
R^2	0.667						
Degrees of freedom	28.254				27.603		

Notes. The semi-parametric estimates are the values from individually estimated models. The degrees for freedom for the semi-parametric models are the average across the 15 estimated models.

Figure 1
LWR with GCV Window Size

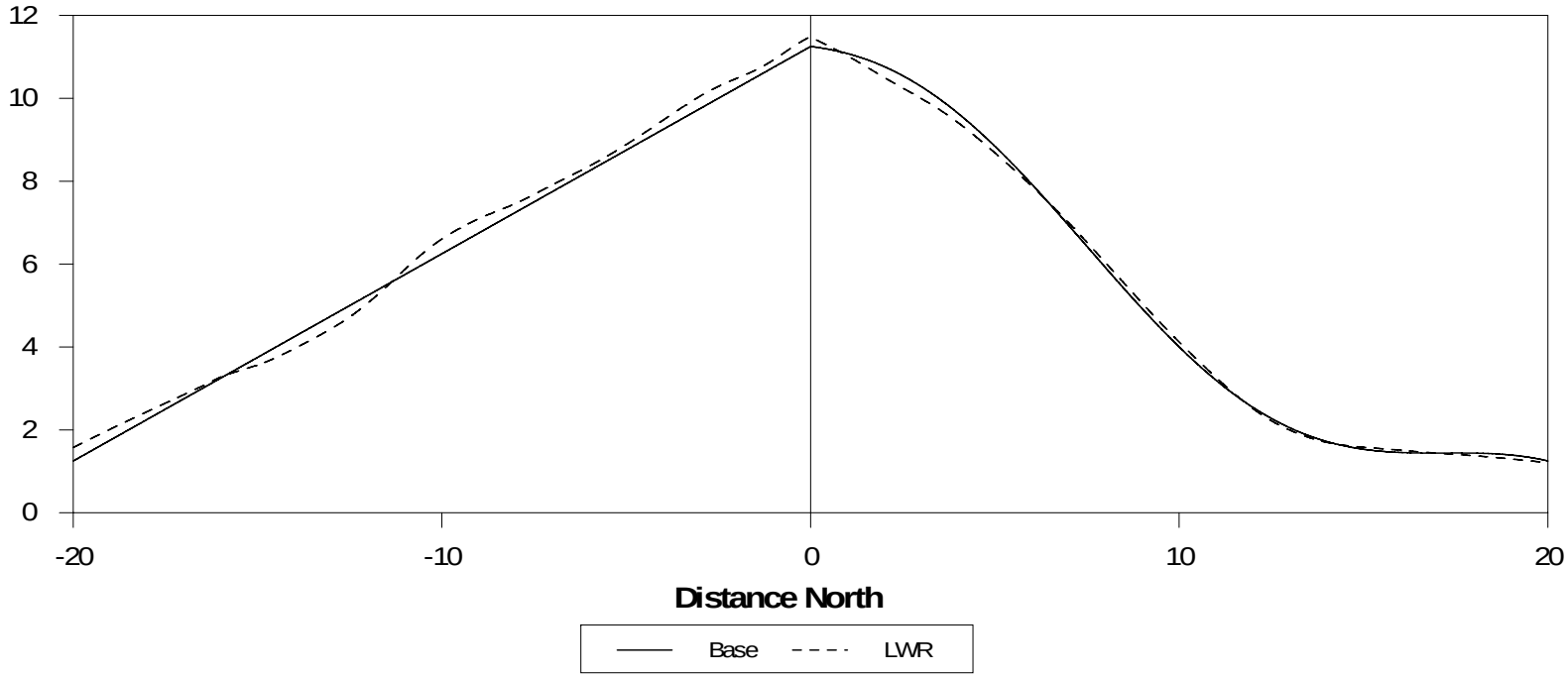

Figure 2
Alternative Kernels

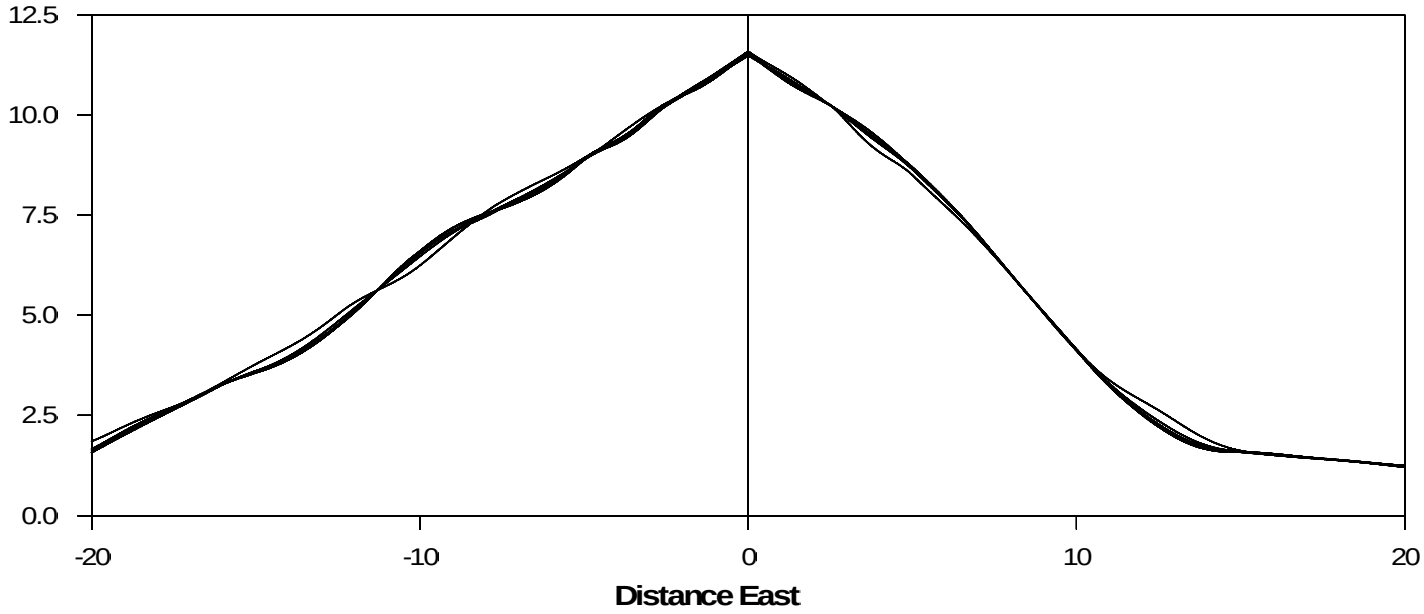

Figure 3
Alternative Window Sizes

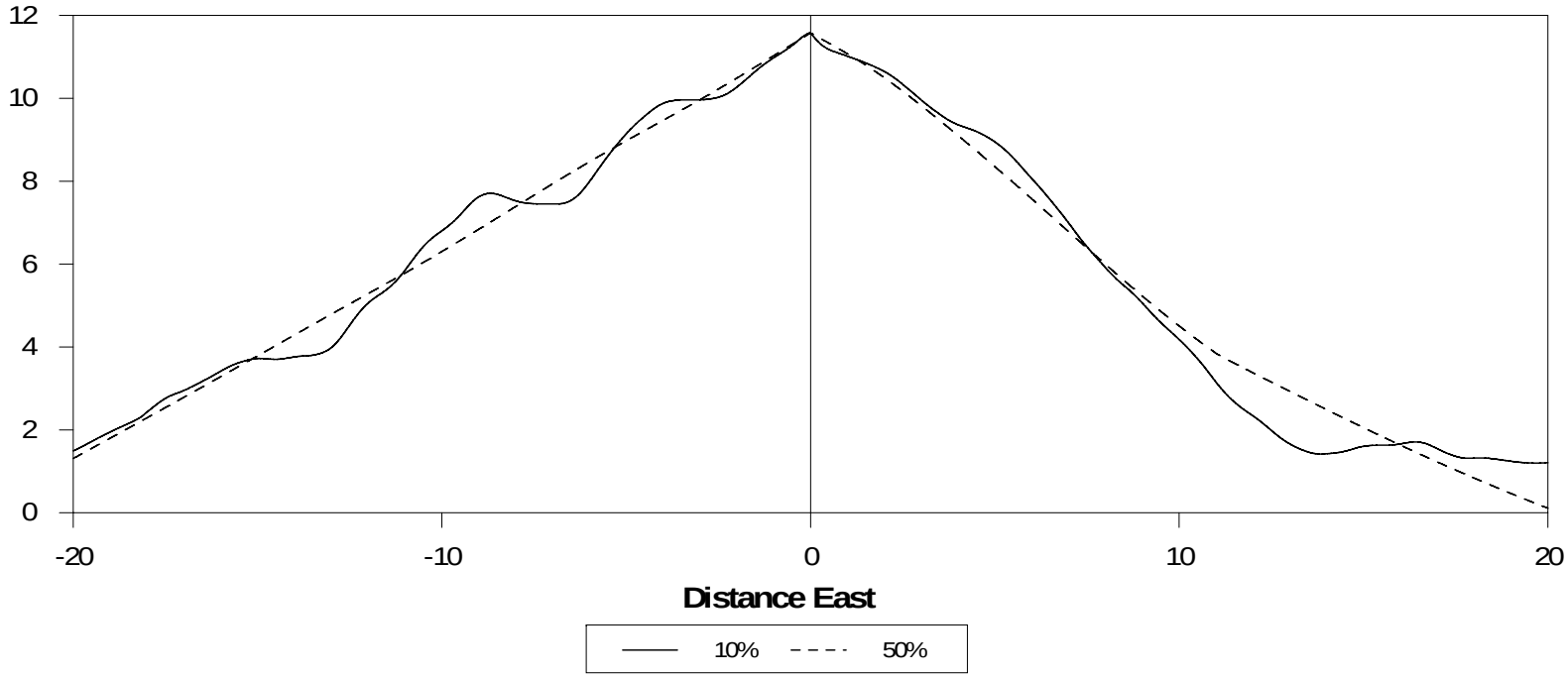

Figure 4
Marginal Effects

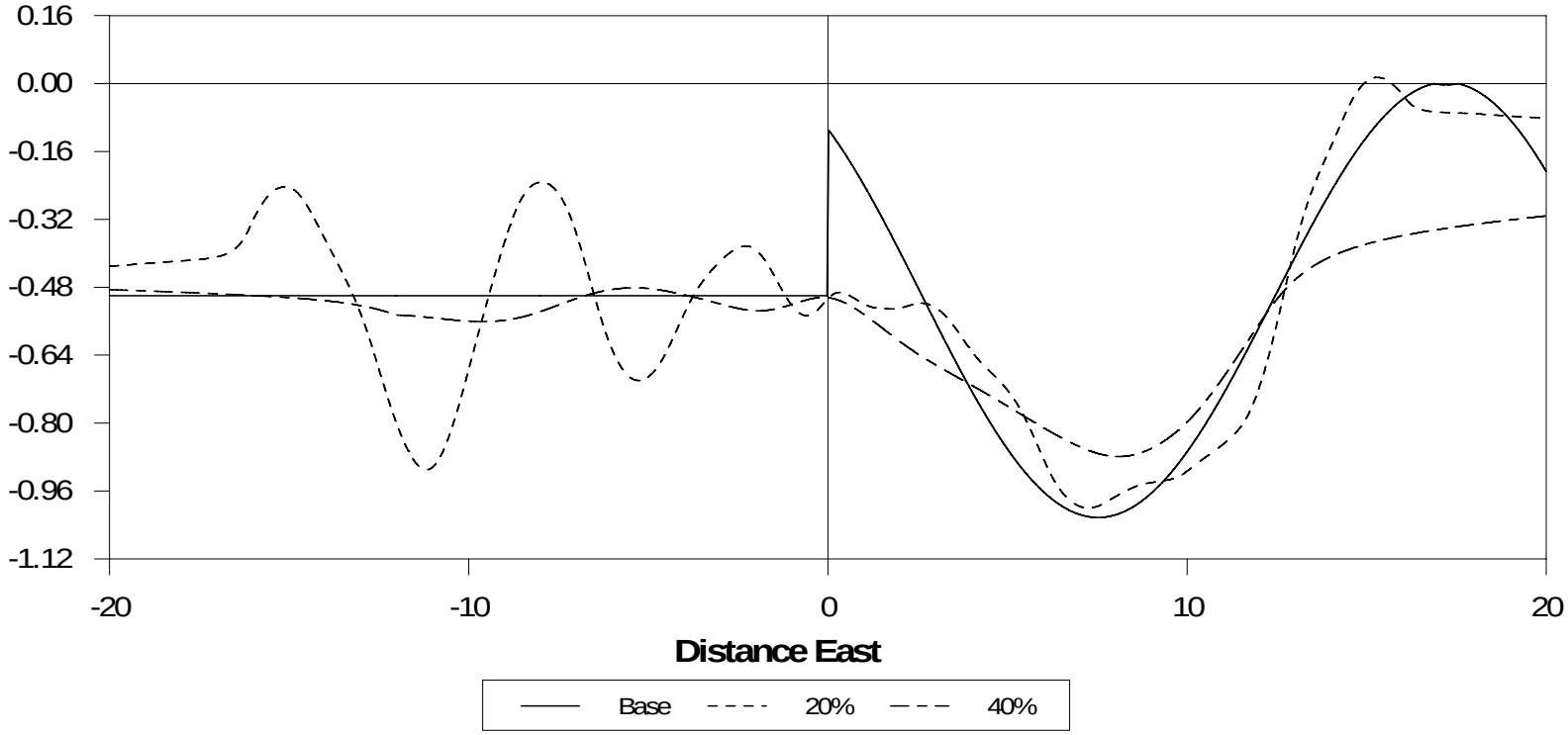

Figure 5

Fixed Effects Model

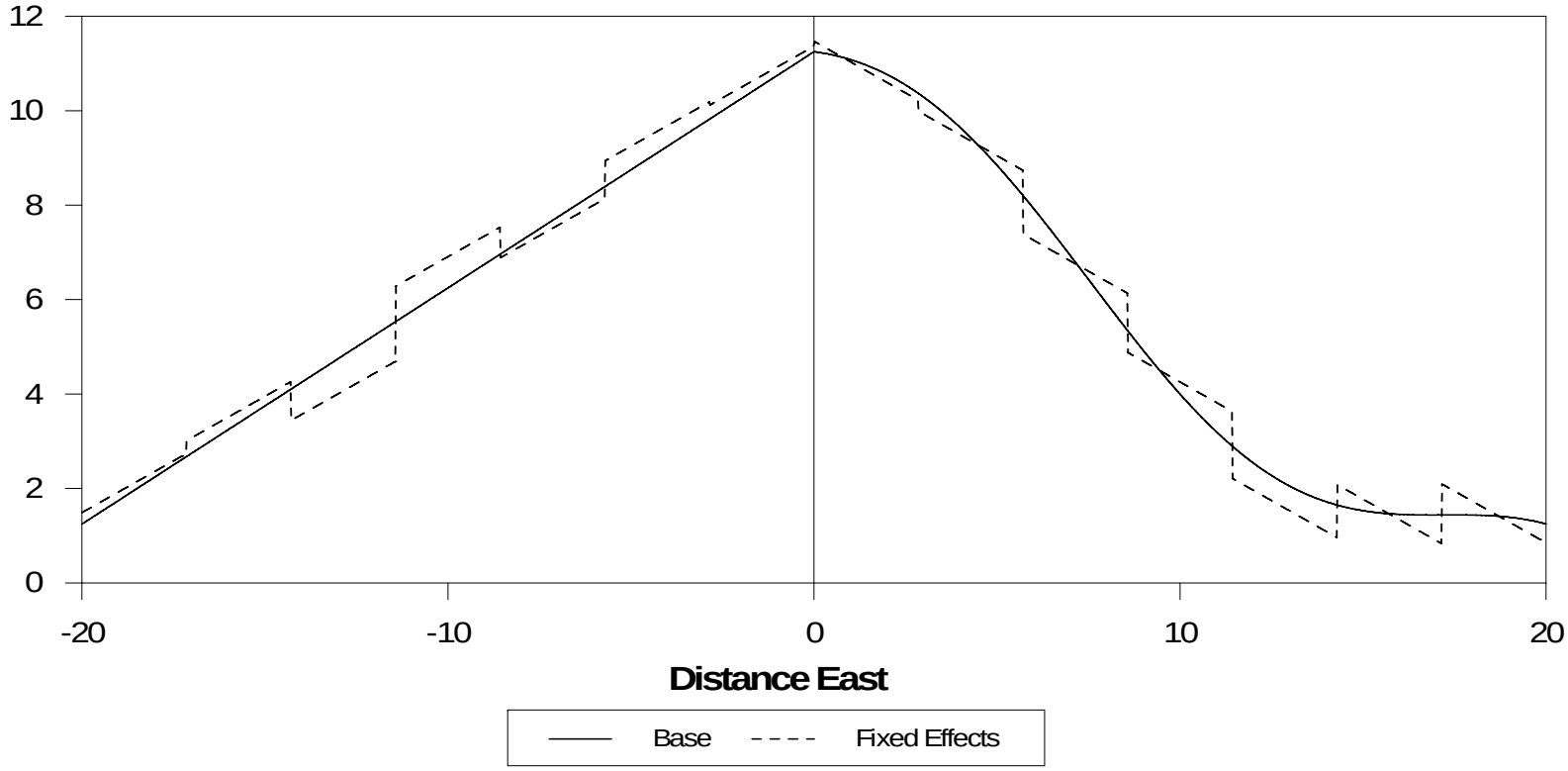

Figure 6
Estimated CPAR Coefficients for Distance from the Nearest EL Stop

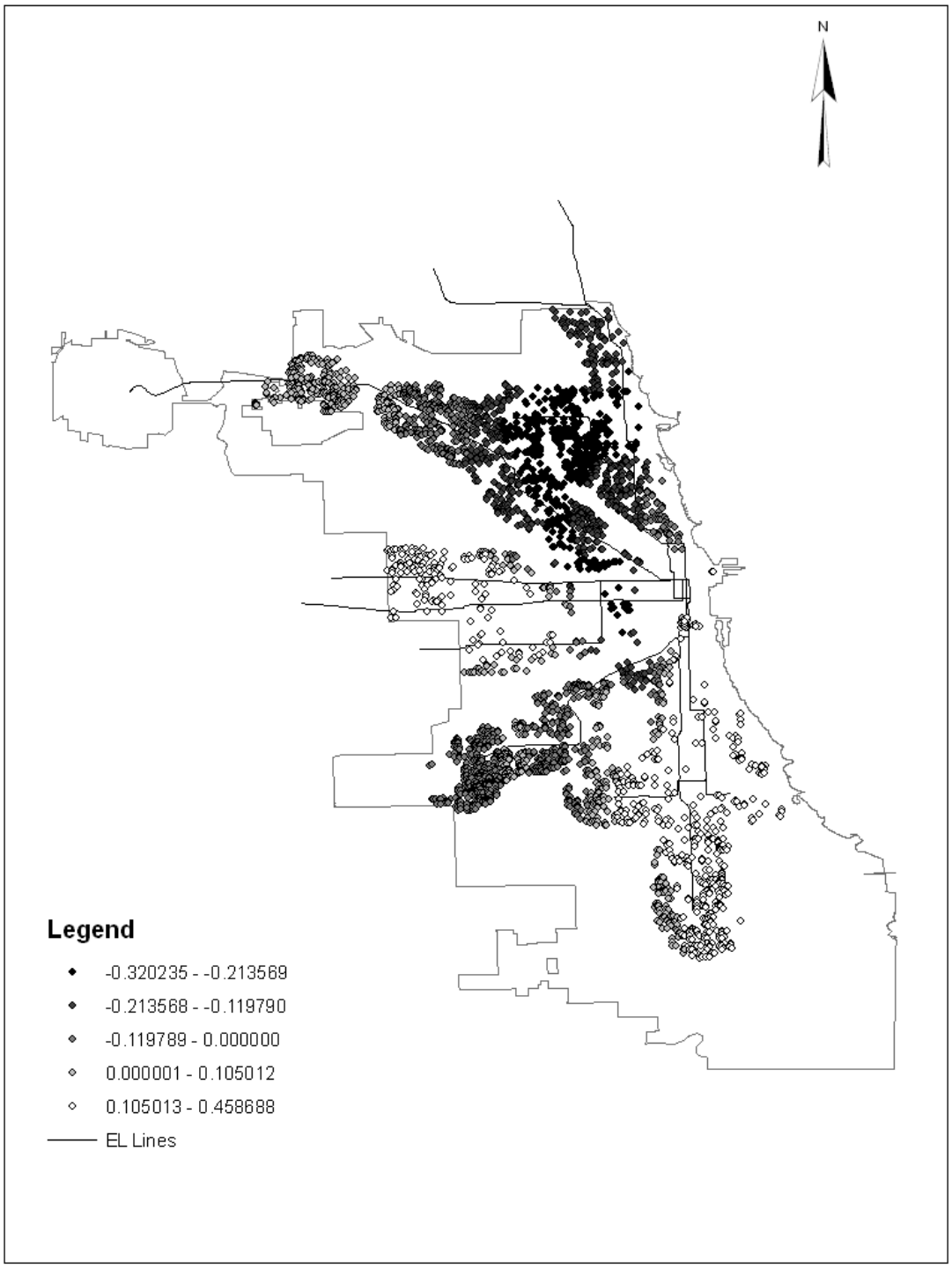

Figure 7
The Distribution of Estimated Coefficients for Distance from the Nearest El Stop