

Revisiting the Empirical Foundations for Measuring
the Capitalization of Access to Transit

Final Report

METRANS Project 08-07

June 2011

Principal Investigator: Christian L. Redfearn

School of Policy, Planning, and Development

University of Southern California

Los Angeles, CA 90089

email: redfearn@usc.edu

office phone: 213-821-1364

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, and California Department of Transportation in the interest of information exchange. The U.S. Government and California Department of Transportation assume no liability for the contents or use thereof. The contents do not necessarily reflect the official views or policies of the State of California or the Department of Transportation. This report does not constitute a standard, specification, or regulation.

Abstract

In METTRANS project number: 04-18, “Transit Investment & the Capitalization of Access into Land Values,” it was shown that the typical statistical tool – “standard hedonic regression” – used for assessing the capitalization of the transit access had potentially fatal flaws for use in informing policy making. Indeed, I showed that a set of commonly-applied hedonic analyses of house prices and transit capitalization produced highly unstable results in an urban context in which distinct housing submarkets existed in close proximity. This finding may explain the existing literature on capitalization of access to passenger rail; this large literature reveals no consistent estimate of access capitalization into surrounding land and property values. This lack of clear guidance is highly problematic for policy makers looking to make sound policy with regard to transit infrastructure investments.

The specific problems have their origins in the “standard” tools of assessing capitalization in the presence more than one housing submarket in the data – a common occurrence in urban areas. The solution I proposed in METTRANS project 04-18 – a more flexible, non-parametric approach – left several questions unanswered. With regards to the technical aspects of the non-parametric approach, the project provided little guidance as to formal hypothesis testing. In light of feedback the previous research generated, it appeared that this weakness could be addressed. The first part of this research project (METTRANS 08-07) formalizes the statistical basis for non-parametric analysis of capitalization; the second would bring it to data. This first step is an essential element of gaining acceptance in a field dominated by the (demonstrably unstable) hedonic regression. The second step was initially proposed to revisit the Los Angeles data used in the previous research. In the end, a new data set from Chicago was used to provide broader empirical support for the findings.

Contents

1	Introduction	6
2	Background and Motivation	7
3	Measuring the Value of Access to Passenger Rail	9
4	Analysis	11
4.1	Monte Carlo Simulations	11
4.2	Nonparametric Regressions	15
5	Conclusion	22
6	Bibliography	23

List of Figures

1	Comparing Standard Hedonics with LWR Models	12
2	Model Comparisons Under Uncertainty - Known Spatial Fixed Effects	13
3	Model Comparisons Under Uncertainty - Unknown Spatial Fixed Effects	13
4	Estimated CPAR Coefficients for Distance from Nearest EL Stop	20

List of Tables

1	Descriptive Statistics	15
2	Linear Regression Results	16
3	CPAR Results (Window Size = 25%)	17
4	CPAR Results (Window Size = 100%)	18

Disclosure

Project was funded in entirety under this contract to California Department of Transportation.

1 Introduction

The goal of this project was to clarify several points regarding the tools used to assess capitalization of access to passenger rail transit. Earlier research funded by METTRANS (04-18), raised significant issues regarding “standard” approaches to measuring the value of access. These estimates are, of course, essential to cost-benefit analysis of rail infrastructure investment. Despite the obvious importance, decades of research on the issue had not returned a robust estimate. The previous research had suggested several lines of inquiry as to the causes of estimate instability in the existing literature. This project was to follow two of these. In particular, the previous research abandoned “standard” hedonic analysis in favor of a more robust non-parametric alternative – locally weighted regression (LWR). The first question with this approach was how to undertake formal hypothesis testing within a model in which parameters estimates varied widely over space. The second question was “does it matter?” – and the project planned a return to data to undertake the formal testing of the null hypothesis that access to passenger rail was not capitalized into surrounding house prices – making use the advances in hypothesis testing developed in the first part. This report details the process and findings in METTRANS-funded research project 08-07.

Measuring the prices of non-traded goods is difficult. One commonly-used approach is to assess the capitalization of an amenity into house price. As they are fixed in space, the price of a house represents the value of the physical attributes of the dwelling and lot, as well as the total value of the amenities and disamenities to which it is exposed. In the previous work, I suggested an alternative (locally-weighted regression), but much was needed to buttress this suggestion. The research presented here, METTRANS project 08-07, extends the previous work in two ways. First, it provides a clear basis for hypothesis testing when using non-parametric analyses. Second, it takes the results on hypothesis testing to new data, examining the benefits/costs associated with proximity to the “EL” stations in Chicago, Illinois. The findings are consistent with intuition and with the forty-some years of literature that finds no clear answer. The basic tool of these analyses – the hedonic analysis – is sensitive to the existence of submarkets. In urban areas where submarkets are common, seemingly benign choices of functional form and sample size can induce profound changes in parameter estimates. These were the primary findings of METTRANS project 04-18. The current research reveals marked spatial patterns in the valuation of access to the EL system. In this case, the population average is misleading; there are areas in which proximity is highly valued, while there are others in which proximity is a nuisance. This, more nuanced result, is robust and a better starting point for discussions of appropriate public policy regarding transit investment.

This approach is likely to have broad implications for a variety of policy questions. Because standard hedonics have been the basis for so much policy analysis, there are likely to be many other areas of urban policy that may be called into question. The imposition of fixed parameters across

space – as is typically assumed – appears not to hold in general; but the fixed parameter approach is widely used anyhow. In contrast to the rigid structure of standard parametric hedonic analysis, nonparametric estimators control for misspecified spatial effects while using highly flexible functional forms. Despite these advantages, nonparametric procedures are still not used extensively for spatial data analysis due to perceived difficulties associated with estimation and hypothesis testing. We demonstrate that nonparametric estimation is feasible for large data sets with many independent variables, offering statistical tests of the added explanatory power of individual covariates as well as tests of model specification. We show, for example, that fixed parameterization of distance to the nearest rapid transit line is a misspecification and that pricing of access to this amenity varies across neighborhoods within Chicago. Furthermore, we demonstrate that nonparametric alternatives to standard hedonics can use significantly fewer degrees of freedom than regressions with neighborhood fixed effects.

2 Background and Motivation

Urban theory is clear on the impact of changes in transportation costs, however empirical evidence of support for the theory is not so decisive. Theoretically, lower transit costs allow households to consume more of other goods with income that would otherwise would have been spent on commuting. In the original urban models, “other goods” referred to land, with lower transit costs leading to a more dispersed urban form. The empirics of measuring the impact of changes in commuting costs are far more complex in a world with heterogeneous agents, a built environment that is expensive to change, lengthy planning horizons, irregular zoning, and an ever-changing metropolitan economy. The lack of clear empirical evidence as to the benefits of transit investment speaks more to the complexity of the problem and lack of appropriate data rather than to the sense that there is, in fact, little benefit. The motivation for this research is to develop a robust method for the assessment of capitalization of access to mass transit. The need for such a tool is clear, with hundreds of millions of dollars of transit investment planned. The policy process requires reliable estimates of how access to these services are valued.

The first step in this research agenda was taken with METTRANS project 04-18, “Transit Investment & the Capitalization of Access into Land Values.” In it, the common-practice tools of hedonic analysis were shown to be unstable – providing significantly different parameter estimates based on sample size and functional form. This distinct lack of robustness is highly problematic for policy makers. As some level, it implies that “answers” can be “chosen” by making seemingly innocuous choices regarding sampling and functional form.

Nonparametric models offer significant advantages for hedonic price function estimation. Functional form flexibility is a common feature of all nonparametric procedures. In addition, nonparametric procedures are readily adaptable to the special requirements of spatial data sets, allowing coefficients to vary over space. 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 spatial 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.

One appeal of the standard hedonic approach is that individual variables can be tested for significance in a very straightforward manner. An analogous mechanism is required of the nonparametric tools. The second technical shortcoming of the previous work is that it does not develop a method for calculating effective degrees of freedom. Again, where the standard approaches make this calculation trivial, the locally weighted regressions appear to have no degrees of freedom – because there is a regression at each data point. This is false, of course, due to the imposition of smoothness conditions, but the calculation of the effective degrees of freedom is an important requisite for acceptance of a non-parametric alternative to the standard hedonic approach to assessing capitalization.

These technical contributions are necessary for this undertaking. They represent significant contributions and have been published in the *Journal of Regional Science*. The larger goal remains the valuation of the access to mass transit. The previous work took a strong stance on the benefits of temporal and spatial variation in the access to rail stations in recovering an unbiased estimate of its value. That is, new stations offered a “clean” identification of capitalized transit services that may be confounded with other amenities/disamenities in cross-sections or studies of existing stations. This implied that, because of its age, much of the existing highway system in Los Angeles was beyond a clean assessment. In keeping with this argument, the previous study analyzed housing sales before and after the development of the Red and Gold Metro Lines.

Where the independent effect of transit access is the desired parameter, this “differences-in-differences” approach is technically correct. However, for many cost-benefit or welfare calculations, the net benefits may be entirely appropriate as inputs. In the proposed research two valuations are then to be undertaken. The first is an analogue to the original work, but with the inclusion of the more formal hypothesis testing and degrees of freedom calculations. These steps should serve to buttress the initial findings of low systematic capitalization of access to rail service in houses around rail stations. The second valuation exercise would take a much broader view of access and measure access to both light rail and highway access throughout the County of Los Angeles. This would allow for an analysis of the factors that influence access valuations as a function of neighborhood characteristics.

Clearly, understanding how markets value access is important. Benefits of access to the light rail system should be captured in the immediate-surrounding areas, with impact a function of proximity.

Any cost-benefit analysis is inadequate in the absence of the indirect benefits that are captured by local land markets. Indeed, transit provision may be substantially misguided if the benefits are mismeasured. Additionally, the cost to society of restrictive zoning or nimby-ism cannot be known with any certainty without an accurate estimate of the value that markets place on transit access.

The research proceeds in two steps: Monte Carlo simulation to infer basic statistical properties and build intuition and non-parametric analysis of the value of access to passenger rail in metropolitan-area Chicago. 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 and 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 non-parametric and semi-parametric procedures can be used to assist in the specification of a hedonic house price function. In this case study, we focus on the valuation of access to an important amenity, Chicago's elevated rapid transit line (the "EL"). In the traditional approach to the valuation of this amenity, neighborhood fixed effects would be employed to control for spatial variation in house prices while the effect of proximity to the EL would be fixed. We show this type of specification to be inappropriate. Not only is the specification rejected statistically, maps of the LWR results reveal systematic patterns in the coefficients suggesting proximity to the EL is not valued equally throughout the city. These patterns have a natural microeconomic foundation: 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 Chicagos 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. In these neighborhoods, the EL is frequently lined by vacant lots and failing businesses, which may lead to the positive correlation between house prices and distance from the EL. Thus, the nonparametric LWR estimates reveal a source of model misspecification that would be far from obvious in a more traditional parametric estimation strategy. Of course, standard parametric hedonic analysis could, in principle, be used to capture the same regularities. In practice, however, local fixed-effect interactions are undertaken in an ad hoc manner with interactions generally tested only for specific variables of interest. The remaining variables are left fixed, a potential misspecification that can lead to biased parameter estimates. The reason for this one-variable-at-a-time approach to spatial effects is clear: generally there are not enough degrees of freedom to interact all the covariates with all the fixed effects. This is the area of perhaps greatest confusion over nonparametric approaches. Paradoxically, the flexibility inherent in LWR comes with greater efficiency, using far fewer degrees of freedom than the fixed effects method an attractive feature in housing research, given the scarcity of sales data.

3 Measuring the Value of Access to Passenger Rail

The project ended up being a collaborative one, with Dan McMillen of the University of Illinois. Dan is the co-editor of *Regional Science and Urban Economics*, one of the top two journals on these issues. My earlier METRANS research (04-18) ended up being published there and Dan and I have been regular correspondents on these issues.

As intended, the project focused on the use of locally-weighted regression to assess the value of rail access. This is a departure from standard practice – a set of norms based on hedonic regression. The resistance to non-parametric analysis is based on four arguments: that nonparametric regressions (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. While the primary goal of this project is to recover robust estimates of the value of access, a secondary goal became to address these four factors in a manner that leaves the reader with a strong sense that today’s researchers have many valid alternatives to traditional parametric hedonic analysis. Rather than turning reflexively to traditional applications of hedonic models, we argued that a large and growing literature on alternative methodologies is sufficiently mature that the data and the empirical question at hand should guide the choice of statistical methodology. The question of the value of access is one in which the reflexive choice of “standard” tools may be the cause of the lack of consensus.

We sketched the structure of several related alternatives to parametric analysis, focusing primarily on locally weighted regression (LWR), which under various pseudonyms has become the most commonly employed nonparametric procedure for analyzing spatial data. We show that three other common alternatives to traditional hedonic regression – kernel regression, conditionally parametric regression, and the so-called “geographically weighted” regression – are special cases of LWR. We then use a combination of Monte Carlo techniques and a representative housing dataset to illustrate how nonparametric estimation procedures can readily be used in formal hypothesis testing and as a means of checking a model specification.

Using Monte Carlo procedures, we showed that LWR uses far fewer degrees of freedom than fully nonparametric estimation while achieving impressive degrees of predictive accuracy and can readily be used to conduct formal hypothesis tests for large datasets. Using a representative dataset of single-family home sales in Chicago for 2000, we showed how nonparametric and semi-parametric procedures can be used to assist in the specification of a hedonic house price function. In the research case study, we focused on the valuation of access to an important amenity, Chicago’s elevated rapid transit line (the EL). In the traditional approach to the valuation of this amenity, neighborhood-fixed effects would be employed to control for spatial variation in house prices while the effect of proximity to the EL would be fixed. We showed this type of specification to be inappropriate. Not only is the specification rejected statistically, maps of the LWR results reveal systematic patterns in the coefficients suggesting proximity to the EL is not valued equally throughout the city. These

patterns have a natural microeconomic foundation: 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. In these neighborhoods, the EL is frequently lined by vacant lots and failing businesses, which may lead to the positive correlation between house prices and distance from the EL. Thus, the nonparametric LWR estimates reveal a source of model misspecification that would be far from obvious in a more traditional parametric estimation strategy.

Of course, standard parametric hedonic analysis could, in principle, be used to capture the same regularities. In practice, however, local fixed-effect interactions are undertaken in an ad hoc manner with interactions generally tested only for specific variables of interest. The remaining variables are left fixed, a potential misspecification that can lead to biased parameter estimates. The reason for this "one-variable-at-a-time" approach to spatial effects is clear: generally there are not enough degrees of freedom to interact all the covariates with all the fixed effects. This is the area of perhaps greatest confusion over nonparametric approaches. Paradoxically, the flexibility inherent in LWR comes with greater efficiency, using far fewer degrees of freedom than the fixed effects method and an attractive feature in housing research, given the scarcity of sales data.

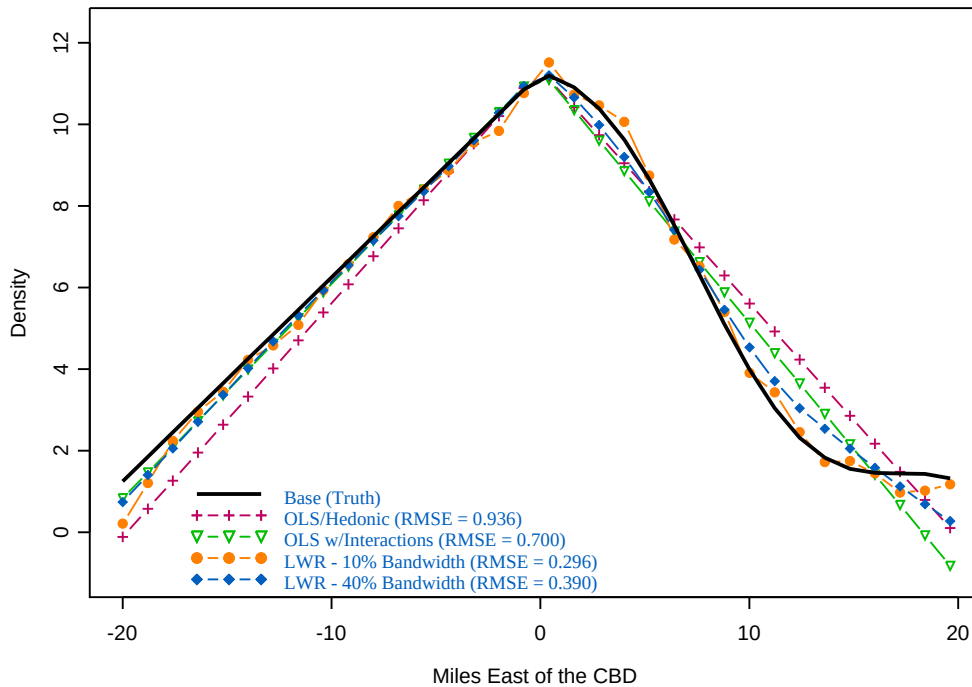
4 Analysis

4.1 Monte Carlo Simulations

In order to provide some insights regarding the benefits of nonparametric regression and contrasts with "standard" versions of hedonic analyses, we developed two sets of Monte Carlo experiments based on a stylized model of an urban area. In the first set of experiments, the dependent variable declines continuously with distance from the city center while the second set includes additional discrete spatial influences. The base function for both sets of experiments looks much like a population density function in a monocentric city in which log-densities decline with distance from the city center – though, for illustrative purposes, it is mildly asymmetric. Around these broad spatial trends, the second experiment appends slightly irregular "neighborhoods," each adding a fixed shift to the base function. The simulations serve two purposes. First, the simpler, no-fixed effects simulation highlights the fundamental differences between "standard" hedonics and "nonparametric" hedonics, i.e., between globally linear models and locally linear models. The second set of simulations is designed to make this same difference clear in the

Figures 1 – 3 plot the base function for the two different base models. The figures also plot the fitted values for a variety of standard OLS and nonparametric specifications. The easier model to fit is the first – with no neighborhood-fixed effects. Figure 1 shows the base function as well as fitted values from a fixed-parameter OLS regression, an OLS regression allowing for different slope

Figure 1: Comparing Standard Hedonics with LWR Models



coefficients east and west of the CBD, and two LWR surfaces using different bandwidths.

It is easy to see that the fixed parameter OLS model performs poorly relative to the more flexible models. Though it captures broad spatial trends, it imposes a single slope parameter on both sides of the city where the slope varies by direction and distance from the CBD. The OLS regression with interactions between east/west dummies and the distance variable performs better. It fits well west of the CBD, but is forced to be linear where the surface to the east of the CBD is nonlinear. The final two lines in Figure 1 show the estimated lines from LWR models with y as the dependent variable; $|x|$ (i.e., distance from the city center) as the explanatory variable; weights based on the geographic distance between each observation j and the target point i , or $|x_i - x_j|$; and two window sizes: 10 percent and 40 percent. Thus, this model can be thought of as a LWR, CPAR, or GWR model. We use a tri-cubic weight function. The estimates closely approximate the true line even though the base model is linear in x . In the 10 percent bandwidth model, the surface moves above and below the true function with greater frequency than the 40 percent model; this is due to smaller samples and greater exposure to outliers. The drawback of the 40 percent model can be seen to the far east of the CBD, where this LWR surface over-smooths.

A more complex underlying surface with irregular fixed effects is used in Figures 2 and 3. In Figure 2, the boundaries of the “neighborhood” fixed effects are known a priori and the models are specified correctly. Conversely, Figure 3 more closely approximates reality in that the true neighborhood boundaries are not known by the researcher when specifying the regression model

Figure 2: Model Comparisons Under Uncertainty - Known Spatial Fixed Effects

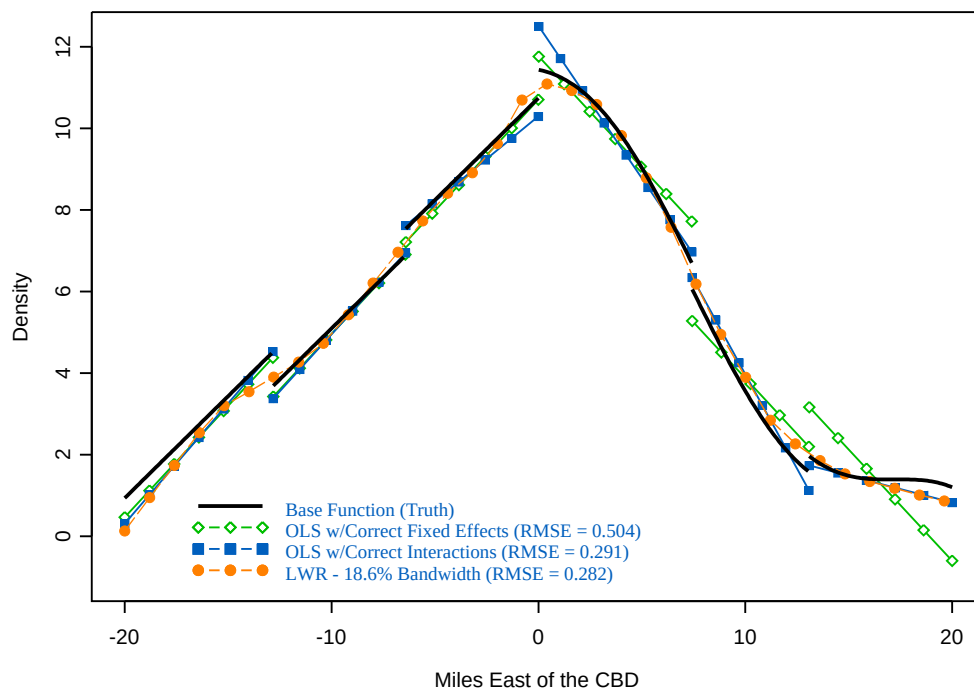
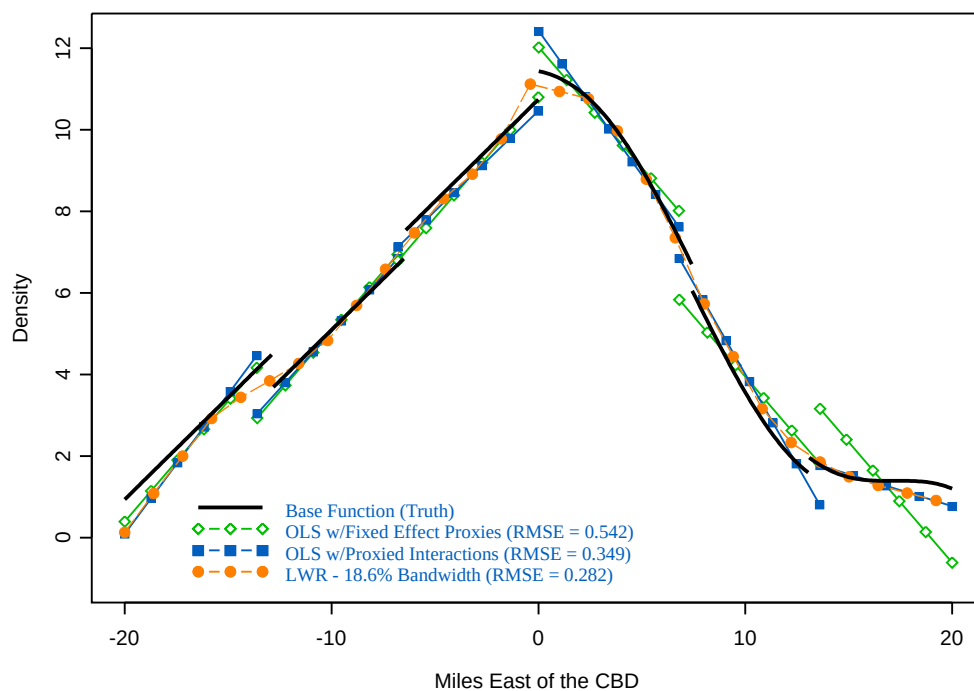


Figure 3: Model Comparisons Under Uncertainty - Unknown Spatial Fixed Effects



– each of the OLS models uses proxies for the underlying fixed effects that are measured with some error. To simulate the effects of mismeasurement of the neighborhood boundaries, the six fixed effects shown in Figure 2 are imprecisely specified in Figure 3. That is, the fixed effects in the estimated models in Figure 3 are set at the actual variables plus or minus as much as a 10 percent error. Thus, for a given boundary x , the boundary used in estimation for Figure 3 is set at $x * (1 + \delta)$, where $1 + \delta$ represents a draw from a uniform distribution with a range of 0.9 to 1.1.

In Figure 2, the OLS regression with a single slope parameter and correctly specified fixed effects is shown with diamonds. As before the linear west side is well approximated, the fixed effects offer the requisite flexibility. On the east side, the nonlinearity again causes problems when a single slope parameter is used. This result should be contrasted with the more complete OLS model that incorporates both correctly specified fixed effects and distance interactions, which allow for flexible estimation of linear slope parameters. Here the linear approximation within neighborhoods works well and the RMSE is essentially the same as that reported for the LWR. It is easy to see the trade-offs. At each step in the underlying function, the LWR makes a smooth transition. And while the flexible OLS model allows for the discrete shift to be captured, the imposed linearity is inappropriate in the eastern half of the urban area. This fully interacted model does as well as the LWR, but achieving this required a priori knowledge of the neighborhood boundaries; no such requirement was needed for the LWR.

Figure 3 demonstrates how the OLS model becomes less accurate when the boundaries are not known precisely in the estimation. Even small random errors in the neighborhood proxies result in relative large losses in RMSE. The RMSE jump by 20 percent in response to only an approximately 5 percent level of mismeasurement of neighborhood areas. In general, the fixed effects used in empirical work are census tracts, Zip codes, or some other administrative boundary. It is unlikely that any of these will capture precisely the spatial extent of a locational amenity or the set of locational amenities. Figure 3 is, thus, more representative of the reality of empirical work in urban areas: amenities are likely to be omitted, administrative boundaries are likely to be noisy proxies for the set of amenities and disamenities influencing house prices or other local outcomes, and the econometrician rarely knows the true structure of the underlying model. In the face of such uncertainty, Figure 3 suggests that a flexible tool like LWR is more likely to allow the data to reveal the underlying surface.

It should be noted that this result is not due to the use of more information as the window size for the LWR in Figures 2 and 3, 0.186, is the value of h that uses 12 degrees of freedom, which is exactly the amount consumed in the estimation of the fully interacted OLS model. This is perhaps the single most important misperception about nonparametric approaches: each parameter estimated is not independent. Instead, smoothness is imposed by the nature of the overlapping samples, which results in fewer degrees of freedom being used in the estimation of nonparametric models one might expect.

Table 1: 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.

4.2 Nonparametric Regressions

In this section, we illustrate the benefits of nonparametric estimation in a typical hedonic price function setting. The dataset includes all sales of single-family homes in 2000 that were within onemile 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 1. The base hedonic specification is Model 1 in Table 2. With the natural logarithm of sales price as the dependent variable and a combination of structural

Table 2: 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
R2	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.

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 percent lower one mile from an EL stop than at the station. This estimate might be used by policy makers to assess the potential benefit of building a new EL line or opening a new stop.

Hedonic estimates can be sensitive to the specification, particularly to the specification of locational effects (Redfearn, 2009). A common method of controlling for local fixed effects is the inclusion of census tract dummy variables. Model 2 of Table 2 adds 466 census tract fixed effects to the base regression. Not surprisingly, the results change substantially. Although the R2 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 of freedom are required for this model, or about 13 percent of the number of observations in the dataset.

In these data, as in practice, some of the geographic units will produce noisy estimates of the fixed effect. This can arise due to the amount of heterogeneity in the local stock relative to the number of sales. In response, combining proximal geographic units may be considered. But, here there is little guide for the appropriateness of these mergers. Small samples do not obviate the requirement that the model be correctly specified; as shown in the Monte Carlo simulations, misspecified fixed-effects models perform significantly worse than more flexible alternatives.

Tables 3 and 4 show CPAR and semi-parametric estimates at two window sizes, 25 percent

Table 3: CPAR Results (Window Size = 25%)

	Nonparametric		Semi-parametric				
	Mean	Std.	F-Test	P-Value	Semi-	Std.	P-Value
	Non-Param.	Std. Dev.			Semi-Param.	Std. Err.	
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
R2	0.825						
Degrees of freedom	164.144				141.024		

Notes. The semi-parametric estimates are the coefficients from individually estimated models in which the listed variable is the sole parametric component of an otherwise fully nonparametric model. The degrees of freedom for the semi-parametric models are the averages across the estimated models.

and 100 percent. All of the models are estimated using a tri-cubic 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.

Table 4: CPAR Results (Window Size = 100%)

	Nonparametric				Semi-parametric		
	Mean Non- Param.	Std. Std. Dev.	F-Test	P-Value	Semi- Semi- Param.	Std. 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
R2	0.667						
Degrees of freedom	28.254				27.603		

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

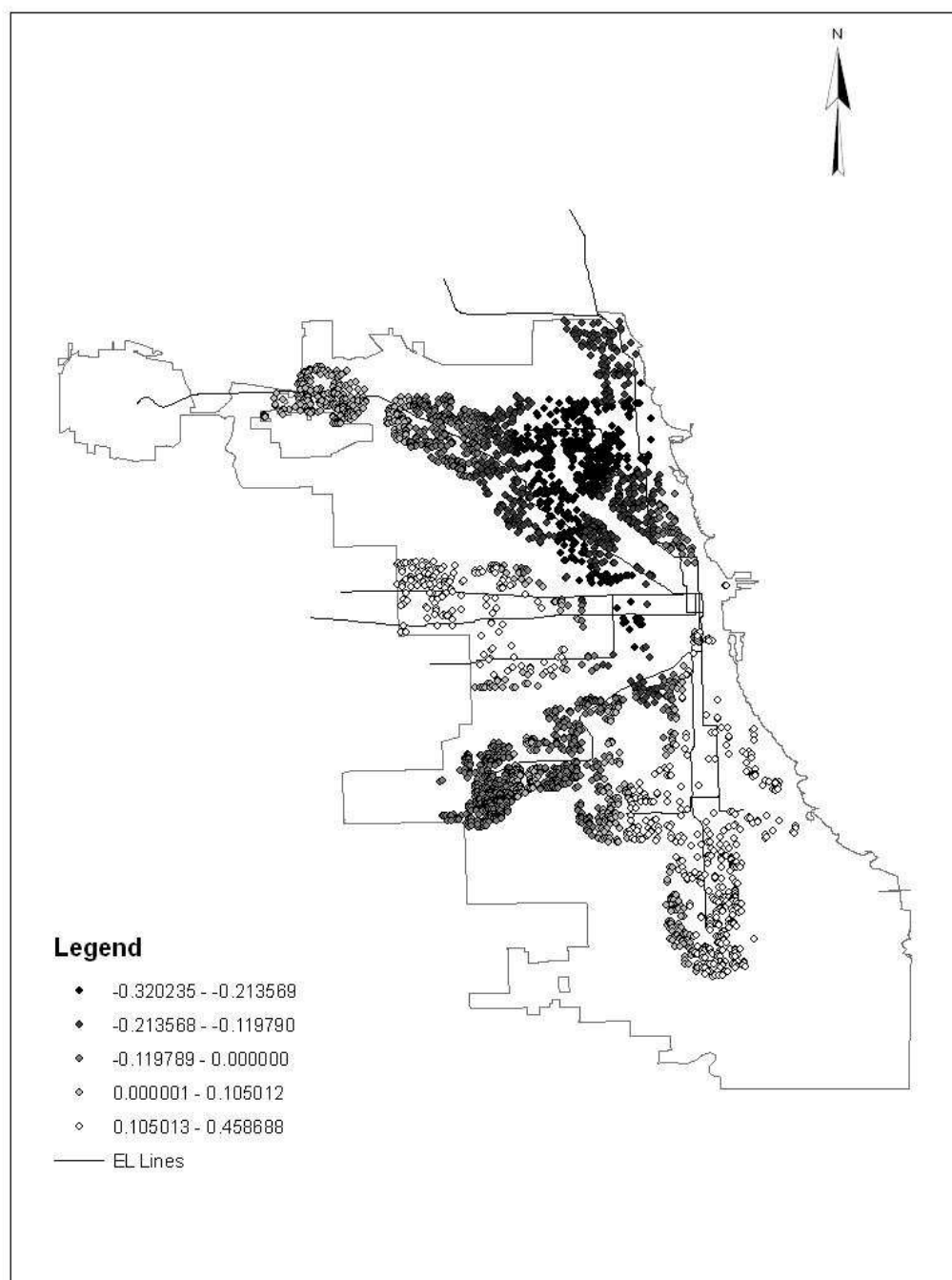
The listed explanatory variable is constrained to enter the model parametrically, while all other variables enter the model in CPAR form with spatially varying coefficients. The tables show (1) the average coefficients from the nonparametric CPAR models and (2) parametric coefficients from the semi-parametric 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 3 and 4 suggest that the coefficients for the structural characteristics are quite stable on average and produce results that are similar to their counterparts in Table 2. For example, the average coefficient for the log of building area is 0.315 across the 3,705 CPAR estimates when the window size is 25 percent, 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 3,705 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 percent in Table 4. The estimates can be compared to the linear regression models of Table 2, in which the base estimated coefficient of 0.403 falls to 0.281 when the model includes census tract fixed effects. The 25 percent window size produces results comparable to the fixed effects specification, while the 100 percent 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 receiving 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. Pagan and Ullah (1999) suggest 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 percent window size is likely preferable to the 25 percent 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 percent window size, the results for our primary variable of interest suggest that prices decline on average by nearly 17 percent 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 semiparametric result suggests that the magnitude of the marginal effect of this variable is lower at 10.9 percent, but still highly significant. The reason for the difference between the two sets of estimates can be seen in Figure 4, which shows the spatial distribution of estimated CPAR coefficients for the smaller, 25 percent window size. The coefficients are clearly negative for most observations north of the city center, meaning that house prices there 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.

Where the standard OLS parameter measures an average effect, the spatial pattern in the

Figure 4: Estimated CPAR Coefficients for Distance from Nearest EL Stop



coefficients offers information about within-group behavior. It is important to recognize that for some policy questions, the average effect is not as useful as understanding effects for particular subsets of the population and could be misleading. For example, 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 access to the EL is not uniformly

valued and 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. Furthermore, nothing in the flexible models precludes the finding of fixed parameters – they are revealed, not imposed.

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 percent 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 percent window, the CPAR model uses far fewer degrees of freedom than the census tracts fixed effects models – 164 v. 483. (Note that this figure of 483 is the number of fixed effects with fixed implicit prices, i.e., without interactions between the fixed effects and other explanatory variables. Interactions use many more degrees of freedom.) The commonly held impression that all nonparametric estimators lead to very high variances in models with more than one or two explanatory variables is mistaken. Many degrees of freedom are saved by taking advantage of the special nature of spatial datasets to include only distance in the kernel weight function.

The second point to note when comparing CPAR and linear results is that the hypothesis tests provide similar 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 variable's overall statistical significance. The standard error for a semi-parametric estimate provides the same information as its counterpart from a linear regression. The difference is that the nonparametric and semi-parametric versions of the model control more effectively than standard linear models for nonlinearities and spatial heterogeneity.

Although our focus has been on hypothesis and specification testing, nonparametric approaches also have significant advantages for purposes of prediction. To evaluate the predictive ability of the models, we hold out a random sample 400 observations and reestimate the simple OLS, census tract fixed effects, and LWR models using the remaining 3,305 data points. We then calculate the RMSE for the 400 holdout observations. The RMSE is 0.287 for the LWR model with a 25 percent window, compared with 0.426 for the OLS model without fixed effects and 0.255 for the model with census tract fixed effects. It is not surprising that the fixed effects model provides the best out-of-sample fit since uses nearly three times as many degrees of freedom as the LWR model. The LWR model much more efficiently achieves nearly the same level of out of sample accuracy as an extremely large fixed effects model.

5 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 dataset, 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 datasets, nearby has a natural geographic interpretation but can also have an attribute-space interpretation, and the conditionally parametric model allows marginal effects to vary smoothly over either definition of 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. The 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 semi-parametric models can be used to construct confidence intervals for the marginal effects of individual explanatory variables.

The literature on alternatives to traditional hedonic analysis is growing, but remains fractured and hard to characterize simply. This has had the effect of impeding the adoption of sensible improvements in the set of statistical tools available to researchers. While we have been frugal in our representation of a large and diverse literature, we have tried to make clear several points regarding hedonic analysis beyond the familiar fixed parametric approach. First, while there are many labels for the alternatives, many share a common architecture and set of modeling elements. Second, they are easy to implement, as they can be recast simply as repeated applications of WLS. Third, rather than being harder to interpret than a single point estimate, the vector of coefficient estimates offer the possibility of informing researchers about nuances in economic relationships beyond average population effects. Our example of Chicago's EL and its capitalization into local house prices produced results that not only rejected the fixed price model as misspecified, but also offered intuitive and interesting patterns of capitalization that could merit further investigation. Our goal was less to prescribe a single best approach and more to motivate empiricists to become familiar with an expanding set of tools that appear likely to help us understand economic behavior in complex urban markets.

6 Bibliography

References

- BAUM-SNOW, N., AND M. E. KAHN (2000): "The Effects of New Public Projects to Expand Urban Rail Transit," *Journal of Public Economics*, 77, 241–263.
- BOLLINGER, C. R., AND K. R. IHLANFELDT (1997): "The Impact of Rapid Rail Transit on Economic Development: The Case of Atlanta's MARTA," *Journal of Urban Economics*, 42(2), 197–204.
- BOWES, D. R., AND K. R. IHLANFELDT (2001): "Identifying the Impacts of Rail Transit Stations on Residential Property Values," *Journal of Urban Economics*, 50, 1–25.
- CERVERO, R., AND M. DUNCAN (2001): "Transit's Value-Added: Effects of Light and Commuter Rail Services on Commercial Land Values," *University of California – Institute of Transportation Studies Working Paper*, pp. 1–19.
- CERVERO, R., AND K.-L. WU (1997): "Polycentrism, Commuting, and Residential Location in the San Francisco Bay Area," *Environment and Planning A*, 29(5), 865–886.

- CLEVELAND, W. S., AND S. J. DEVLIN (1988): “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting,” *Journal of the American Statistical Association*, 83(403), 596–610.
- GIBBONS, S., AND S. MACHIN (2005): “Valuing Rail Access Using Transport Innovations,” *Journal of Urban Economics*, 57(1), 148–169.
- KAHN, M. E. (2001): “City Quality-of-Life Dynamics: Measuring the Costs of Growth,” *Journal of Real Estate Finance and Economics*, 22(2–3), 339–352.
- McMILLEN, D. P. (2003): “Spatial Autocorrelation or Model Misspecification?,” *International Regional Science Review*, 26(2), 208–217.
- (2004): “Employment Densities, Spatial Autocorrelation, and Subcenters in Large Metropolitan Areas,” *Journal of Regional Science*, 44(2), 225–244.
- McMILLEN, D. P., AND J. McDONALD (2004): “Reaction of House Prices to a New Rapid Transit Line: Chicago’s Midway Line, 1983-1999,” *Real Estate Economics*, 32(3), 463–486.
- MEESE, R., AND N. WALLACE (1991): “Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices,” *American Real Estate and Urban Economic Association Journal*, 19(3), 308–332.
- SMITH, L. B., AND M. H. HO (1996): “The Relative Price Differential between Higher and Lower Price Homes,” *Journal of Housing Economics*, 5, 1–17.