Spatial Variation in House Prices and the Opening of Chicago's Orange Line

Daniel McMillen
University of Illinois
Department of Economics
1407 W. Gregory Dr.
Urbana, IL 61801
mcmillen@illinois.edu

April 2018

Abstract

The recent literature on the effect of public transit lines on property values has relied primarily on difference in difference estimation approaches. The standard difference in difference approach of comparing sales prices before and after the opening of a new transit line or station is not well suited to this analysis because prices typically rise well in advance of the new service and there is no clear treatment or control area. Moreover, estimated treatment effects are likely to vary spatially and are not necessarily uniform across the full spectrum of home prices. An analysis of home prices before and after the opening of Chicago's Orange Line from the Loop to Midway Airport suggests that house prices began to rise near the new line about two years before the opening. Estimated appreciation rates are highest for relatively low priced homes, and there is signification spatial variation within the area close to the new line.

1. Introduction

The large literature on the effects of transit lines on house prices began with straightforward hedonic regressions of house prices on distance to stations, typically using cross sectional data sets. Representative examples of the hedonic approach include Dewees (1976), Bajic (1983), and Voith (1993). Although the hedonic approach can establish whether prices are higher near transit stations, the results are likely to be biased by missing variables that themselves are correlated with transit access. Studies such as McMillen and McDonald (2004); Chatman, Kim, and Tulach (2012); Kim and Lahr (2013); and Pilgrim and West (2018) use repeat sales as a way to control for the effects of missing variables when analyzing appreciation rates near new rail lines. Similarly, Baum-Snow and Kahn (2000) use Census data to estimate the effect of changes in distance to a station after a new transit opens on the change in property values

Another set of studies use difference in difference (DD) approaches to estimate the effect of new transit lines on house prices. Examples include Gibbons and Machin (2005); Billings (2011); Zheng et. al. (2016); Diao, Leonard, and Sing (2017); Wagner, Komarek, and Martin (2017); and Pilgrim and West (2018). These studies compare appreciation rates using house sales before and after the opening of new stations for a set of treatment and control areas. Although the difference in differences approach is widely used in urban economics (Baum-Snow and Ferreira, 2015), it generally is not well suited to spatial data sets where the appropriate definition of control and treatment locations may be unclear and can vary over locations. For example, a host of studies have analyzed the effects of enterprise zones by comparing outcomes for variables such as house prices and unemployment rates for enterprise zones to arbitrarily defined control areas. The enterprise zone is a clearly defined area; the control locations (e.g., contiguous zones, a ring around the enterprise zones, or "similar" areas) may or may not be suitable alternatives, and authors such as Hanson and Rohlin (2013) have found that estimated treatment effects are sensitive to the definition of the control area.

The problem is still worse for cases such as new transit lines where the extent of the treatment area is also not clearly known. For example, Gibbons and Machin (2005) define their treatment area as any location in London that became closer to a transit stop after a new Tube line was opened, while the control locations were all other areas in London. Although this definition might be suitable in a place like London where nearly all locations are well served by the Tube, it does not work well for American cities where the new, closer station is still too far away to be considered a viable alternative. Most studies have instead drawn rings around the new sites, with close rings being identified as "treatment" areas and distant rings called "control" areas. A typical study then varies the widths to see if the results are stable. For example, Pilgrim and West (2018) define treatment as being within ½ mile of a new transit line, while the control groups are (1) homes from the rest of South Minneapolis, (2) homes within a ½ - 1 mile ring, and (3) any home in Minneapolis that is more than ½ mile from the new line. Diao, Leonard, and Sing (2017) compare appreciation rates for properties that are within 600 meters (the treatment area) to properties that are between 600 and 1600 meters of a new transit line in Singapore. They then estimate separate effects for properties in 0-200, 200-400, and 400-600 rings. These approaches are typical of the DD approaches used in analyses of spatial data for what is, in fact, an underlying continuous measure of distance from a treatment site such as a new transit station.

My objective in this paper is to demonstrate how a standard DD approach can be combined with additional statistical procedures to provide a more comprehensive picture of the change in prices associated with a new transit line. The data set is an expanded version of the one used in McMillen and McDonald (2004) to analyze the effects of Chicago's Orange Line, which opened service between the Loop and Midway Airport in 1993. Whereas McMillen and McDonald analyzed repeat sales, I follow the approach taken by recent studies and use a DD approach for the full sample of sales from 1988-2006. The treatment area is defined to be the area within 0.5 miles of the Orange Line (and also closer to an Orange Line station than to a station on another line). The control area is comparable: the area within 0.5 miles of one of the other lines on Chicago elevated train system (known locally as the "EL", although some portions are underground).

I generalize the standard DD analysis by allowing the "treatment" effects to vary by year, including years before the opening of the line. The results confirm McMillen and McDonald's (2004) finding that the effect of the Orange Line was capitalized into nearby house prices at least two years prior to its opening, and was especially high in 1992, the year prior to the opening of the line. The results are not sensitive to the width of the ring drawn around the new line to define the treatment area, apart from a small range of highly industrial properties close to the line where the effect is insignificant. A combination of quantile regression and locally weight regression analysis suggests that these average effects are far from uniform: appreciation rates are much higher for relatively low-priced properties and for properties near the stations closest to downtown Chicago.

The variation in the estimated appreciation rates is far from trivial. The range in estimated appreciation rates for 1988-1992 is 36.0% - 61.1%. The range is still greater when the estimates are allowed to vary by quantile. For the 0.10 quantile, the variation in estimated appreciation rates across locations is 16.9% - 72.7%, with an average of 37.9% Comparable values for the median and the 0.90 quantile are 29.8% - 58.9% and 31.0% - 56.8%. The results imply a much more complex set of effects of the new line than is implied by an analysis that is focused more narrowly on a single, average effect.

2. Empirical Approaches

The base empirical model is a slight variation of the standard difference in differences specification:

$$lnP_i = x_i\beta + \delta_t + \lambda_c + \theta_t Orange_i + u_i \qquad (1)$$

where P_i is the sale price of property *i*, which is located in census block group c and which sold at time *t*. X_i is a vector of structural characteristics for the property. In the empirical application, the structural characteristics include building area, lot size, rooms, bedrooms, bathrooms, age at the time of sale, and variables indicating the presence of a basement, central air conditioning, brick construction, and a garage with 1 or 2+ spaces. The fixed effects δ_t and λ_c control for time trends and characteristics of the census block group that do not change over time. $Orange_i$ is a variable indicating that property i is located closer to a station on Chicago's Orange Line than to other elevated (EL) train stations. In contrast to the standard DD specification, equation (1) allows the coefficient on $Orange_i$ to vary over time, which is critical in this application because new rail lines take a long time to build and their effects on the housing market are clearly anticipated well in advance of their opening, (e.g., McMillen and McDonald, 2004; Diao, Leonard, and Sing, 2017; an McDonald and Osuji, 1995; and Pilgrim and West, 2018).

Equation (1) separates the sample into two groups – the treatment area that is directly affected by the opening of the new EL line (*Orange* = 1), and a control area where prices are not affected by the new line (*Orange* = 0). Although this specification is in contrast with the early literature as well as more recent studies such as Baum-Snow and Kahn (2000) and McMillen and McDonald (2004) that estimate the rate of decline of house prices with distance from new transit stations, it follows the trend of DD studies that have recently dominated the literature. The goal of the DD approach represented by equation (1) is to estimate an average treatment effect rather than to estimate the spatial variation of a new station on house prices. The key question is how to specify the treatment and control areas reasonably accurately.

For the base specification, I define the treatment group as any sale that is within 0.5 miles (about 800 meters) of a new station on the Orange Line and that is closer to an Orange Line station than to a station on another EL lines.[1] Control group observations include sales that are within 0.5 miles of a station on one of Chicago's other EL lines. Alternatives include distances ranging from 0.25 to 1.5 in increments of 0.25 miles.

All of these specification impose a single, common effect of the new EL line on all sales that are within an arbitrary distance. Although an average is interesting, it may conceal significant variation in the effects over both location and type of housing. In a series of papers, I have advocated using locally weighted regression (LWR) procedures to allow for spatial variation in estimated coefficients (e.g., McMillen, 1996; McMillen and Redfearn, 2010). The LWR approach is based on an approach developed by Cleveland and Devlin (1988) and Cleveland, Devlin, and Grosse (1988), and was introduced to the urban literature by Meese and Wallace (1992). The LWR approach involves estimating weighted least squares regressions for a series of target points, with more weight placed on observations that are closer to the targets. Following McMillen (1996) and most subsequent urban LWR approaches, I base the measure of distance on the simple straight-

[1] I also delete 47 sales that are in Chicago's Near South Side community area because the observations are close to Orange Line shares stops with the Red Line in this neighborhood.

line distance of each observation from the target point, an approach that has come to be known as "geographically weighted regression." The results can then be interpolated to other points in the data set.

Since the objective of the LWR analysis is to determine whether there is significant variation in appreciation rates in the area served by the Orange Line, I use only the Orange Line sample in the locally weighted regressions. The base specification includes the structural coefficients and the time fixed effects. The LWR version of the model is:

$$lnP_i = x_i\beta(z_i) + \delta_t(z_i) + u_i \qquad (2)$$

where z_i represents the geographic coordinates for observation i. The LWR approach makes it possible to determine whether the new rail line has a larger effect on sales prices in some locations than in others.[2] The results are readily displayed in maps of the sample area.

The effect of a new rail line may vary by price as well as by location. For example, it may turn out that relatively low-priced homes appreciate more than high-priced homes if higher-income households continue to commute by car rather than public transportation following the opening of the new transit line. Quantile regression allows the coefficients to vary across the distribution of sales prices. Following the notation in Koenker (2005), let τ represent the quantile and define $\rho_\tau(u_{ict}) = u\big(\tau - I(u_{ict} < 0)\big)$. The standard conditional quantile regression approach involves finding the values of the coefficients that minimize $\sum_i \rho_\tau(lnP_i - x_i\beta - \delta_t - \lambda_c + \theta_t Orange_i)$.

[2] In the empirical application, I use a tri-cube kernel for the weight function, and a window of 25% of the nearest observations. Since my focus is on the treatment area, I estimate equation (2) using only the treatment census block group centroids as target points, which leads to 388 target points. In practice, this approach involves estimating 388 regressions using the 25% of the sample observations that are closest to the target centroid and weights that decline with straight-line distance from the centroid.

The conditional quantile approach is readily adapted to the locally weighted regression approach (McMillen, 2015). At a target location z, the weight given to observation i is given by the kernel weight function $w_i(z)$, and the quantile estimates for quantile τ at location z is obtained by minimizing a weighted version of the standard quantile objective function:

$$\sum_i w_i(z)\rho_\tau(lnP_i - x_i\beta(z_i) - \delta_t(z_i) - \lambda_c(z_i) + \theta_t(z_i)Orange_i) \quad (3)$$

As with a standard linear regression model, the estimates can be interpolated from the targets to other locations in the sample. The locally weighted approach makes it possible to determine whether the effect of the new transit line varies across locations and by price.

3. Data

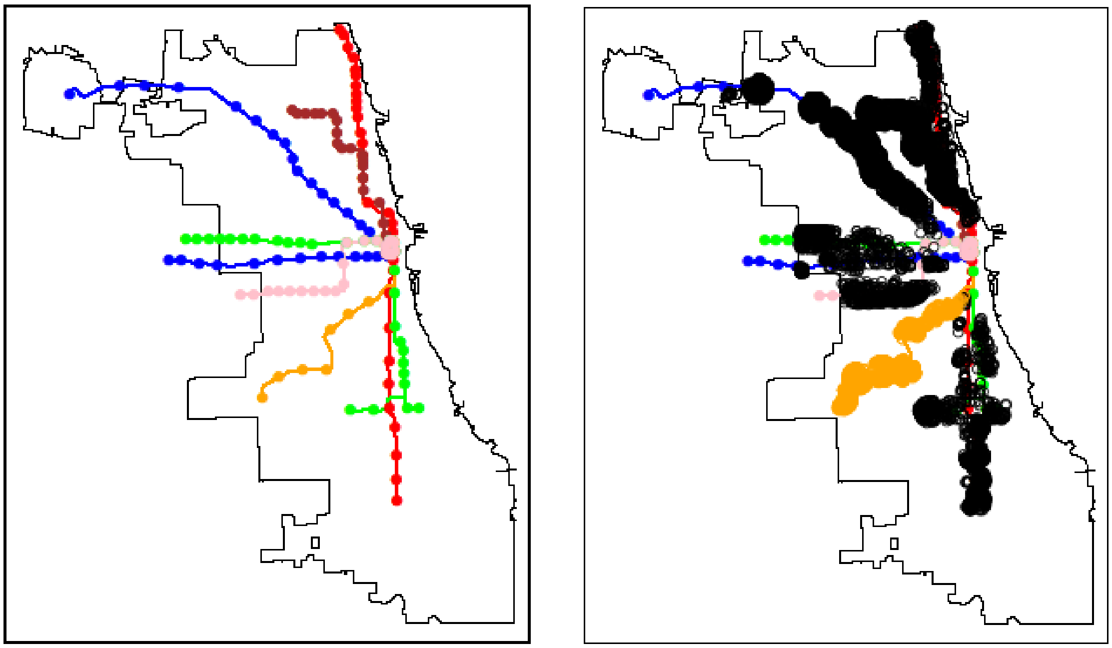
The data set employed in this paper is the same as in McMillen and McDonald (2004), with three significant exceptions: I use data for a larger region, do not limit the sample to repeat sales, and now have data for sales from 1992 as well as for other years around the opening of the new EL line. Construction on the new line began in 1985 and it opened on October 31, 1993. The data set for this paper includes all sales of single-family homes that sold in Chicago between 1988 and 1996. The primary data source is the Illinois Department of Revenue, which collects data on sales to conduct their annual assessment ratio studies. Data on structural characteristics come from the Cook County Assessor's Office.

Descriptive statistics are presented in Table 1. The three time intervals are based on the finding in McMillen and McDonald (2004) that the new transit line had a strong effect on house prices near the new line by 1991. Since the line opened near the end of 1993, the three time periods correspond approximately to a time before the new line was reflected significant prices (1988-1990), the time when prices increased in anticipated of the opening of the line (1991-1993), and the time after the line opened (1994-1995). The left panel of Figure 1 shows the location of Chicago of the color-coded EL lines, and the right panel adds the locations of the sales that are within 0.5 miles of the lines.

Table 1: Descriptive Statistics (13,795 sales)

Variable	1988-90	1991-93	1994-96	1988-90	1991-93	1994-96
Log of Sale Price	11.020	11.372	11.466	11.398	11.603	11.788
Log Building Area	6.985	7.039	7.015	7.268	7.303	7.319
Log Lot Size	8.148	8.165	8.175	8.151	8.147	8.126
Rooms	5.208	5.348	5.260	6.251	6.344	6.426
Bedrooms	2.780	2.855	2.785	3.260	3.297	3.307
Basement	0.646	0.630	0.593	0.706	0.675	0.687
Bathrooms	1.268	1.299	1.285	1.563	1.643	1.688
Air Conditioning	0.098	0.100	0.101	0.194	0.215	0.244
Attic	0.408	0.405	0.344	0.473	0.430	0.426
Brick	0.530	0.547	0.516	0.440	0.461	0.467
1-Car Garage	0.278	0.259	0.281	0.264	0.280	0.284
2+ Car Garage	0.492	0.495	0.462	0.452	0.441	0.449
Age / 10	6.171	6.088	6.448	7.266	7.356	7.450
Distance to EL Line	2.505	2.559	0.286	0.249	0.243	0.245
Number of Observations	817	881	987	3,365	3,903	3,842

Figure 1: EL Lines and Sales Locations



4. Difference in Difference Results

The base DD results are presented in Table 2. The estimated coefficients for the structural characteristics are generally as expected. Prices are higher for newer, larger homes that are made of brick and on larger lots, and have more rooms, bathrooms, and a garage. Sales prices rose steadily throughout the sample period, with an increase of nearly 50% between 1988 and 1996. Homes located near the future Orange Line sold for lower prices than homes along existing EL lines in the base year of 1988.

Consistent with the repeat sales results of McMillen and McDonald (2004), the Orange Line first began to have a significant effect on house prices in 1991, two years prior to the opening of the line. The biggest effect of the new line was in 1992, the year prior to its opening. By 1996, the price premium associated with the new line drops to insignificance, and prices on near the Orange Line are estimated to be approximately the same as prices near other lines.

Table 2: Base Regression Results

Variable	Estimate	Std. Err.	T-Value	Variable	Estimate	Std. Err.	T-Value
Log Building Area	0.278	0.021	0.000	1991 Sale	0.215	0.017	0.000
Log Lot Size	0.358	0.020	0.000	1992 Sale	0.292	0.020	0.000
Rooms	0.010	0.004	0.012	1993 Sale	0.335	0.018	0.000
Bedrooms	0.008	0.007	0.292	1994 Sale	0.379	0.018	0.000
Basement	0.014	0.010	0.168	1995 Sale	0.401	0.017	0.000
Bathrooms	0.047	0.008	0.000	1996 Sale	0.482	0.019	0.000
Air Conditioning	0.010	0.010	0.319	Orange x 1989	0.014	0.025	0.561
Attic	-0.025	0.009	0.006	Orange x 1990	0.033	0.030	0.265
Brick	0.043	0.011	0.000	Orange x 1991	0.129	0.032	0.000
1-Car Garage	0.071	0.012	0.000	Orange x 1992	0.155	0.033	0.000
2+ Car Garage	0.076	0.011	0.000	Orange x 1993	0.098	0.029	0.001
Age/10	-0.033	0.003	0.000	Orange x 1994	0.096	0.030	0.001
Near Orange Line	-0.131	0.026	0.000	Orange x 1995	0.122	0.031	0.000
1989 Sale	0.100	0.016	0.000	Orange x 1996	0.023	0.032	0.463
1990 Sale	0.183	0.017	0.000				

Notes: The dependent variable is the natural logarithm of sale price. The number of observations is 13,795. The regression includes 388 fixed effects for census block group, and standard errors are clustered at the block group level. The R^2 is 0.807.

Table 3 presents results for variations in the maximum distance from an EL station. The estimated coefficients for Orange x Year are not statistically significant when the ring is small (0.25 miles is equivalent to about two blocks in Chicago's regular grid system). Figure 1 shows why this result is not surprising: the Orange Line goes through an area of light industry because it was built along existing rail rights-of-way. The estimated treatment effects become significant when the maximum distance is increased to 0.50 miles, and it remains both stable and statistically significant when the distance is increased. For all distances, the estimated appreciation rate near the Orange Line stations is not statistically different from the rates near other EL stations for 1989 and 1990, and the anticipated effects are first evident in 1991 for all distances.

Table 3: Variation of Maximum Distance from EL Stations

	0.25	0.50	0.75	1.00	1.25	1.50
Near Orange Line	0.346 (0.065)	-0.131 (0.026)	-0.119 (0.076)	-0.103 (0.066)	-0.036 (0.047)	-0.040 (0.038)
Orange x 1989	-0.078 (0.044)	0.014 (0.025)	-0.014 (0.018)	-0.021 (0.016)	-0.013 (0.014)	-0.008 (0.012)
Orange x 1990	-0.035 (0.055)	0.033 (0.030)	0.032 (0.020)	0.022 (0.017)	0.026 (0.014)	0.025 (0.013)
Orange x 1991	0.033 (0.059)	0.129 (0.032)	0.105 (0.022)	0.093 (0.018)	0.081 (0.016)	0.078 (0.015)
Orange x 1992	-0.012 (0.080)	0.155 (0.033)	0.128 (0.025)	0.118 (0.021)	0.116 (0.019)	0.117 (0.019)
Orange x 1993	0.025 (0.059)	0.098 (0.029)	0.095 (0.021)	0.094 (0.019)	0.093 (0.017)	0.085 (0.017)
Orange x 1994	0.042 (0.046)	0.096 (0.030)	0.102 (0.022)	0.095 (0.020)	0.099 (0.018)	0.094 (0.017)
Orange x 1995	0.036 (0.048)	0.122 (0.031)	0.111 (0.023)	0.098 (0.020)	0.096 (0.018)	0.096 (0.017)
Orange x 1996	-0.033 (0.075)	0.023 (0.032)	0.048 (0.025)	0.058 (0.022)	0.061 (0.019)	0.063 (0.018)
Number of Observations	3,948	13,795	24,133	33,260	41,232	49,516
Number of Fixed Effects	225	388	511	572	636	660
R^2	0.834	0.807	0.792	0.783	0.782	0.781

Notes: The dependent variable is the natural logarithm of sale price. Clustered standard errors are presented in parentheses below the estimated coefficients.

5. Local Spatial Variation

The base DD specification leads to an estimate of the average treatment effect for each year from 1989 – 1996. In fact, there is substantial variation in the neighborhoods served by the Orange Line. While much of the line runs through formerly industrial areas, the northeastern end includes a highly residential area that is also well served by other EL lines.[3] In general, areas north of the Orange Line are more likely to be industrial, while areas south of the line are apt to be residential. The last station at the southwest end of the Orange Line serves Midway Airport, which itself is in the middle of a residential area. McMillen and McDonald (2004) report that average weekday ridership ranged from 1,253 at the Ashland Ave. station to 8,531 at the Midway station in 2000. The range in average ridership rose slightly to 1,629 – 8,947, again with Ashland Ave. and Midway being the extremes. Given the range of neighborhood conditions and ridership, it is reasonable to expect that the effect of the Orange Line may differ spatially.

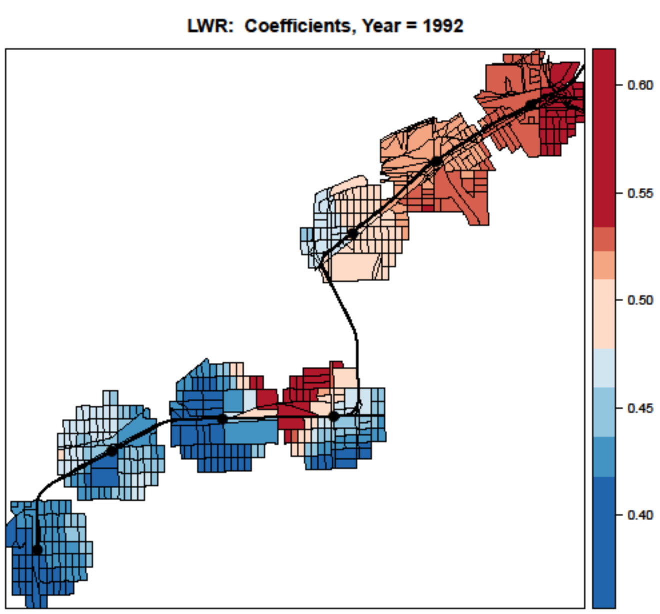
I estimate a set of locally weighted regressions to allow the coefficients of the base regression to vary in the are served by the Orange Line. The target points are the 388 census block group centroids represented in the Orange Line sample area. I then calculate the straight-line from every property in the data set to each target point. Let d_i represent the distance from observation i to a target point, and since the window size is set at 25%, let d^* be the 25th percentile of all of these distances. The weight provided by observation i in estimating the locally weighted regression at the target point is given by the tri-cube function: $k(d_i) = (1 - (d_i/d^*)^3)^3$ for $d_i \leq d^*$ and $d_i = 0$ otherwise. Using matrix notation to re-write equation (1) as $lnP = W\gamma$, the LWR

[3] From southwest to northeast, the order of the stations shown in the figures is Midway, Pulaski, Kedzie, Western, 35th/Archer, Ashland, Halsted, Roosevelt. The Roosevelt St. station is not included in this analysis because it is also a stop for the Red and Green lines. The data used for ridership in 2017 are drawn from the Chicago Transit Authority's Annual Ridership Report for 2017: https://www.transitchicago.com/assets/1/6/2017_CTA_Annual_Ridership_Report.pdf.

estimates are $\hat{\gamma}(z) = (\sum_i (k(d_i) w_i w_i')^{-1})(\sum_i (k(d_i) w_i lnP_i)^{-1})$, where w_i is a column vector holding the elements of W for observation i. Fixed effects for census block groups are omitted from the list of explanatory variables because locally weighted regressions take account of spatial effects by allowing the coefficients to vary smoothly over space rather than confining spatial variation to discrete shifts at block group boundaries in the intercept alone.

The estimated results for the Orange x 1992 variable are shown in Figure 2. The results are representative of the spatial variation for other years also. The results imply that the rate of appreciation for 1988 - 1992 varies from 36.0% to 61.1%. Appreciation rates are highest at the Ashland Ave. and Halsted St. stations, which are the two stops to the extreme northeast of Figure 2. Estimated appreciation rates are also high on the northwest side of the Western Ave. station, which is in the middle of the figure. Estimated rates are lower at the Midway end of the line. The results do not contradict the DD estimates; instead, they provide interesting new information.

Figure 2: LWR Coefficients for Orange x 1992



6. Quantile Regression Results

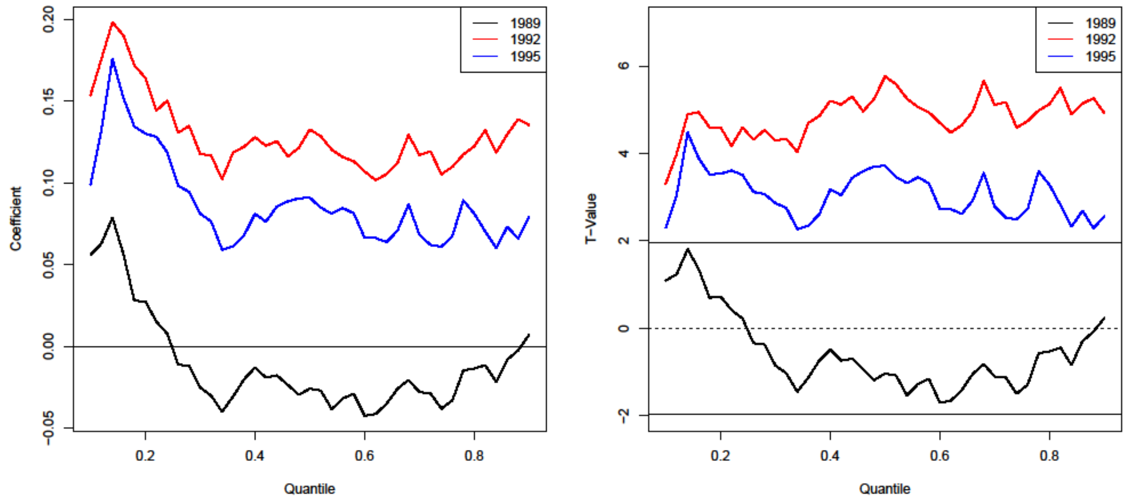
Table 4 shows the results of quantile regression results for three quantiles, τ = 0.10, 0.50, and 0.90. The regressions are estimated using sales from areas near all the EL lines, and the results are nearly directly comparable to the results shown in Table 2. However, as estimating quantile regressions with a large number of fixed effects is cumbersome, I employ a two-stage procedure proposed by Canay (2011) that reduces the dimension of the problem. The first stage is the basic linear regression model of equation (1), which includes controls for both census block groups and the year of sale. The first-stage regression provides estimates of the 388 census block group fixed effects, λ_c. Under the assumption that these effects do not vary across quantiles, the second stage is a standard quantile regression of $lnP - D\hat{\alpha}$ on the remaining variables in equation (1).

Several interesting patterns emerge from Table 4 even though the results are shown for only three quantiles. First, note that the initial discount associated with being closer to the Orange Line than to other lines is larger at -0.188 for τ = 0.90 than for either the median (-0.110) or τ = 0.10 (where the coefficient for *Orange* is -0.051.) This pattern suggests that changing the value of *Orange* from 0 to 1 is associated with a leftward shift in the distribution of log sales prices, with a larger shift at the right side of the distribution. Second, note that the coefficients for *Orange* x Year are statistically insignificant for 1989 and 1990, but they generally are significant for 1991 – 1996. Finally, the coefficients for both τ = 0.10 and τ = 0.90 tend to be larger in magnitude than at the median, although the pattern is not uniform.

Table 4: Quantile Regression Results

	Quantile = 0.10		Quantile = 0.50		Quantile = 0.90	
Variable	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept	5.787	0.183	6.338	0.093	6.986	0.145
Log Building Area	0.230	0.028	0.288	0.012	0.295	0.016
Log Lot Size	0.430	0.013	0.355	0.007	0.294	0.014
Rooms	0.020	0.006	0.002	0.003	0.006	0.004
Bedrooms	0.010	0.010	0.010	0.005	0.010	0.007
Basement	0.064	0.013	0.002	0.007	-0.036	0.009
Bathrooms	0.021	0.011	0.048	0.006	0.079	0.009
Air Conditioning	0.011	0.013	-0.005	0.007	0.006	0.009
Attic	-0.005	0.011	-0.024	0.005	-0.029	0.008
Brick	0.007	0.013	0.029	0.006	0.064	0.008
1-Car Garage	0.142	0.016	0.048	0.007	0.022	0.009
2+ Car Garage	0.126	0.015	0.058	0.007	0.031	0.009
Age / 10	-0.063	0.002	-0.029	0.001	-0.012	0.001
Orange Line	-0.051	0.035	-0.110	0.019	-0.188	0.023
1989 Sale	0.117	0.034	0.121	0.013	0.067	0.020
1990 Sale	0.273	0.029	0.198	0.013	0.115	0.020
1991 Sale	0.267	0.032	0.224	0.012	0.146	0.021
1992 Sale	0.330	0.030	0.300	0.012	0.234	0.019
1993 Sale	0.389	0.030	0.331	0.013	0.272	0.021
1994 Sale	0.453	0.028	0.371	0.012	0.309	0.020
1995 Sale	0.468	0.029	0.388	0.013	0.354	0.021
1996 Sale	0.529	0.028	0.459	0.013	0.430	0.021
Orange x 1989	0.056	0.051	-0.026	0.025	0.007	0.031
Orange x 1990	-0.046	0.055	-0.002	0.025	0.057	0.031
Orange x 1991	0.104	0.057	0.102	0.027	0.125	0.029
Orange x 1992	0.154	0.046	0.133	0.023	0.135	0.027
Orange x 1993	0.069	0.045	0.094	0.025	0.107	0.029
Orange x 1994	0.065	0.043	0.065	0.024	0.081	0.028
Orange x 1995	0.099	0.043	0.091	0.024	0.080	0.031
Orange x 1995	0.006	0.049	0.040	0.024	0.023	0.030

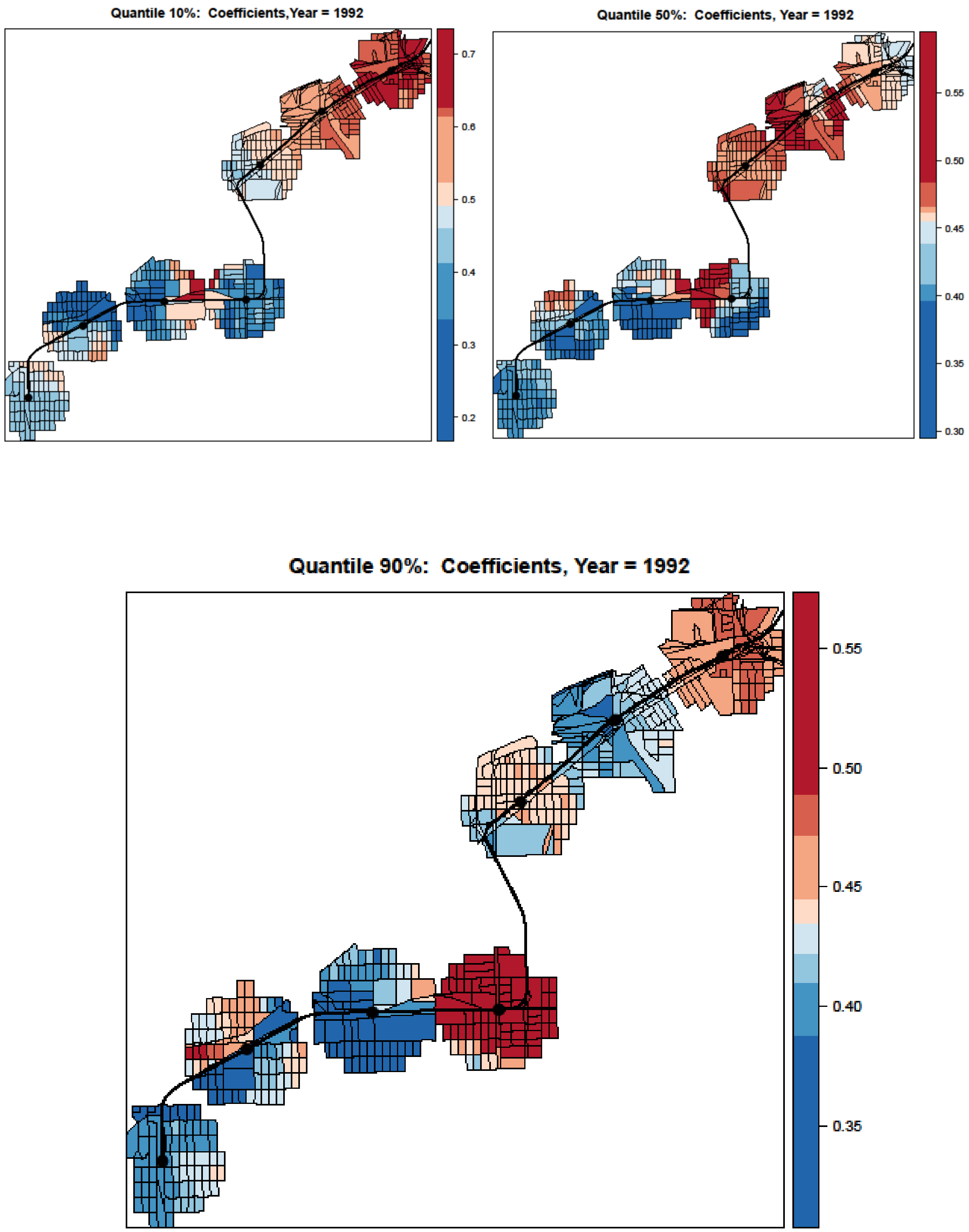
Figure 3: Coefficients and T-Values for Orange x Year by Quantile



Selecting only three sets of quantile regression results can be misleading because the coefficient estimates are not necessarily smooth across quantiles. Figure 1 displays the coefficients and t-values for *Orange* x 1989, *Orange* x 1992, and *Orange* x 1995 across quantiles ranging from 0.04 to 0.96. The left panel shows clearly that appreciation rates are higher at lower quantiles, and the estimates are highly significant in 1992 and 1995. This result implies that the rightward shift in the price distribution was higher in the low-price end of the distribution.

Figure 4 shows the results for the 1992 dummy variable for locally weighted versions of the quantile regressions for $\tau = 0.10$, 0.50, and 0.90. As was the case for the locally weighted regressions, I limit the analysis to the observations that are closer to a station on the Orange Line than to other stations because the objective is to display the spatial variation of appreciation rates within the area served by the Orange Line. I again use a tri-cube kernel with a 25% window, and I omit the census block group fixed effects.

Figure 4: Locally Weighted Quantile Results for Year = 1992



The estimates are more variable for τ = 0.10 than for the other two quantiles, with estimates ranging from 0.169 to 0.727, to a range of 0.298 – 0.589 for τ = 0.50 and 0.310 – 0.568 for τ = 0.90. As was the case for the locally weighted regression estimates, the estimated appreciation rates are highest for the stations on the northeast end of the map, which is closer to downtown Chicago. Appreciation rates are also high for the Western Ave. station in the middle of the map, and are particularly high relative to the other quantiles for τ = 0.90.

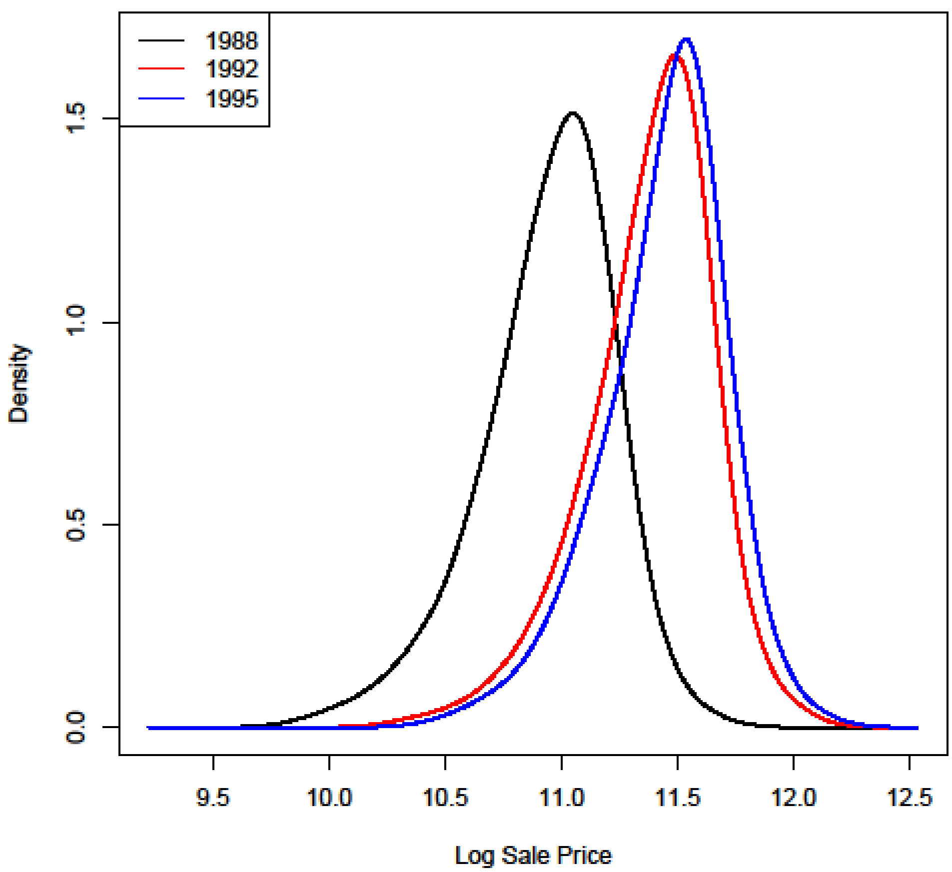
Although it is common to present coefficient estimates at a small number of quantiles, estimating the quantile regressions at many values of τ makes it possible to show how the entire distribution of log sales prices changes when the values change for one of the explanatory variables. The approach is discussed in detail in McMillen (2015). Consider the case of the locally weighted quantile regressions. The estimated log sale price for observation *i* and for quantile τ is

$$x_i\hat{\beta}(z_i,\tau) + \hat{\delta}_{89}(z_i,\tau) + \cdots + \hat{\delta}_{96}(z_i,\tau) \quad (4)$$

Given the actual values of *x*, the predicted log sale price for observation year in 1988 is simply $x_i\hat{\beta}(z_i,\tau)$, and the estimate for, e.g., 1992 is just $x_i\hat{\beta}(z_i,\tau) + \hat{\delta}_{92}(z_i,\tau)$. With *n* observations, these expressions imply *n* predicted values of the log sale price for 1988 and *n* predicted values for 1992, all at quantile τ. Having estimated equation (4) for *Q* quantiles, equation (4) implies *nQ* predicted values for these two years (or any other year). Kernel density functions then can be used for the *nQ* estimates to show how the entire distribution of log sales prices is predicted to have change from 1988 to 1992.

Figure 5 shows the results of these calculations for 1988, 1992, and 1995 after estimating locally weighted quantile regression for τ = 0.04, 0.06, …, 0.96. The distribution of log sales prices shifts well to the right between 1989 and 1992, the latter being the year when the new Orange Line is estimated to have the largest effect on prices. The distribution shifts further to the right in 1995. The increase in the height of the functions over time implies a significant decline in the overall variation in log sales prices. Although it is hard to tell directly from Figure 5, the pattern of the variation by quantile in the coefficients for Year = 1992 implies that the reduction in variation comes about by a greater increase in the left side of the distribution than at the upper end of the distribution.

Figure 5: Predicted Log Sale Price Densities



7. Conclusion

Estimating the effect of a new transit line on property values is representative of the situations where difference in difference estimation approaches have been used routinely in urban economics. Researchers tend to focus on the assumption of common parallel trends for treatment and control areas when attempting to verify the validity of the approach, an assumption that is virtually impossible to meet in housing markets. Any event that is anticipated prior to the formal announcement is likely to be at least partially capitalized into house prices, and house prices do not necessarily appreciate uniformly across neighborhoods even within an urban area. An additional assumption – that there are clearly defined treatment and control areas – is sidestepped by varying the size of rings around a treatment site, often for both the treatment and control areas. This approach is in contrast to the typical approach taken in the earlier urban literature which attempted to measure how prices varied continuously from the treatment site, recognizing that when "treatment" is a discrete site like a rail station, not all sites within a zone receive the same level of treatment.

This paper revisits the McMillen and McDonald (2004) analysis of the opening of Chicago's Orange Line in 1993 to show how locally weighted and quantile regression methods can be used to extend the standard DD analysis to show how prices vary within treatment areas over space and across the price distribution. While McMillen and McDonald focused on repeat sales from the area within 1.5 miles of the Orange Line, I estimate locally weighted versions of the standard DD model using all sales close to both the Orange Line and others lines on Chicago's EL system. As was the case with the earlier study, the results show that the new line began to be capitalized into house prices at least two years prior to its opening, with the largest effect coming in 1992, the year prior to the opening.

But there is significant variation in the magnitude of this effect within the area close to the new line. Locally weighted versions of the DD model imply that the rate of appreciation for 1988 - 1992 varies from 36.0% to 61.1%, with especially high rates around stations closer to downtown Chicago. Quantile regressions imply higher rates of appreciation and more variability in appreciation rates for relatively low-priced homes. Together, these results suggest a much more complex set of patterns than is implied by the simple average treatment effect of a DD model.

References

Bajic, Vladimir, "The Effect of a New Subway Line on Housing Prices in Metropolitan Toronto," *Urban Studies* 20 (1983), 147-158.

Baum-Snow, Nathaniel and Fernando Ferreira, "Causal Inference in Urban and Regional Econoimcs," in Gilles Duranton, J. Vernon Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics* Vol. 5A (2015), Elsevier, Amsterdam, 3-68.

Baum-Snow, Nathaniel and Matthew E. Kahn, "The Effects of New Public Projects to Expand Urban Rail Transit," *Journal of Public Economics* 77 (2000), 241-263.

Billings, Stephen B., "Estimating the Value of a New Transit Option*," Regional Science and Urban Economics* 41 (2011), 525-236.

Canay, Ivan A., "A Simple Approach to Quantile Regression for Panel Data," *Econometrics Journal* 14 (2011), 368-386.

Chatman, Daniel G., Nicolas K. Tulach, and Kyeongsu Kim, "Evaluation the Economics Impacts of Light Rail by Measuring Home Appreciation," *Urban Studies* 49 (2012), 467-487.

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

Cleveland, William S., Susan J. Devlin, and Eric Grosse, "Regression by Local Fitting: Methods, Properties, and Computational Algorithms," *Journal of Econometrics* 37 (1988), 87-114.

Dewees, D.N., "The Effect of a Subway on Residential Property Values in Toronto," *Journal of Urban Economics* 3 (1976), 357-369.

Diao, Mi, Delon Leonard, and Tien Foo Sing, "Spatial-Difference-in-Differences Models for Impact of New Mass Transit Line on Private Housing Values," *Regional Science and Urban Economics* 67 (2017), 64-77.

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

Hanson, Andrew and Shawn Rohlin, "Do Spatially Targeted Redevelopment Programs Spillover?," *Regional Science and Urban Economics* 43 (2013), 86-100.

Kim, Kyeongsu Kim and Michael L. Lahr, "The Impact of Hudson-Bergen Light Rail on Residential Property Appreciation," *Papers in Regional Science* 93 (2014), S79-S97.

Koenker, Roger. *Quantile Regression*. New York: Cambridge University Press, 2005.

McDonald, John F. and Clifford Osuji, "The Effect of Anticipated Transportation Improvement on Residential Land Values," *Regional Science and Urban Economics* 25 (1995), 261-278.

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

McMillen, Daniel P., "Conditionally Parametric Quantile Regression for Spatial Data: An Analysis of Land Values in Early Nineteenth Century Chicago," *Regional Science and Urban Economics* 55 (2015), 28-38.

McMillen, Daniel P. and John F. McDonald, "Reaction of House Prices to a New Rapid Transit Line: Chicago's Midway Line, 1983-1999," *Real Estate Economics* 32 (2004), 463-486.

McMillen, Daniel P. and Christian Redfearn, "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions," *Journal of Regional Science* 50 (2010), 712-733.

Pilgrim, Clemens A. and Sarah E. West, "Fading Premiums: The Effect of Light Rail on Residential Property Values in Minneapolis, Minnesota," *Regional Science and Urban Economics* 69 (2018), 1-10.

Voith, Richard, "Changing Capitalization of CBD-Oriented Transportation Systems: Evidence from Philadelphia," *Journal of Urban Economics* 30, 360-372.

Wagner, Gary A., Timothy Komarek, and Julia Martin, "Is the Light Rail 'Tide' Lifting Property Values? Evidence from Hampton Roads, VA," *Regional Science and Urban Economics* 65 (2017), 25-37.

Zheng, Siqi, Yangfei Xu, Xiaonan Zhang, and Rui Wang, "Transit Development, Consumer Amenities, and Home Values: Evidence from Beijing's Subway Neighborhoods," *Journal of Housing Economics* 33 (2016), 22-33.