Economic Quarterly—Volume 95, Number 1—Winter 2009—Pages 53–74

Semiparametric Estimation of Land Price Gradients Using Large Data Sets

Kevin A. Bryan and Pierre-Daniel G. Sarte

Traditional urban theory typically predicts land values that form a smooth and convex surface centered at a central business district (CBD) (Mills 1972 and Fujita 1989). The fact that land values are highest near the city center reflects a trade off between commuting costs and agglomeration externalities at the CBD. As distance from the city center increases, so do commuting costs for workers employed at the CBD. Agglomeration effects, however, such as knowledge sharing or decreased shipment costs from a common port, are highest near the CBD. In equilibrium, therefore, the price of residential land tends to be bid up most forcefully close to the city center where commuting costs are lowest. In empirical work, the shape of the land price surface is often estimated using a parametric regression that includes a measure of Euclidian distance from the CBD or a polynomial function of location data. The parameter associated with distance from the CBD, then, captures the rate at which land prices decline as one moves away from the city center and toward the rural outskirts. Though this is a straightforward method to obtain estimates of the rate of price decline, parametric methods can be misleading for two reasons.

First, as noted by Seyfried (1963) among others, cities are "not a featureless plain." Bodies of water, mountains, and geography more generally all distort the land price surface by influencing potential commuting patterns. Second, and more importantly, there is growing evidence, both theoretical and empirical, that the monocentric city of Mills (1972), for example, is

▬ We wish to thank Nadezhda Malysheva, Yash Mehra, and Ray Owens for their comments. We also thank Sarah Watt for helpful research assistance. The views expressed in this article do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System. All errors are our own. E-mail: pierre.sarte@rich.frb.org.

being replaced by the polycentric city, where employment subcenters lead to land price gradients of a form that may be difficult to uncover parametrically. Anas, Arnott, and Small (1998) survey this literature, while Redfearn (2007) provides an example of the employment density surface in Los Angeles. A parametric model of such a city may smooth over important employment subcenters and high-price suburbs. As such, nonparametric estimates of the land price surface allow for a more robust description of the data.

Estimation of land gradients using nonparametric or semiparametric methods is somewhat involved relative to parametric regressions. In part, to economize on computations, early work in this area has tended to use only vacant lot sales (Colwell and Munneke 2003), but the sparseness of that data can lead to overly smooth price gradients. Furthermore, vacant lot sales are not as informative when considering land prices outside of dense urban cores, as there may be large areas without any nearby sale during the period studied. In contrast, the number of residential house sales in a given period can be substantial. This article, therefore, reviews a method for constructing land price gradients using a potentially large set of housing sales data. Drawing on work by Yatchew (1997) and Yatchew and No (2001), we estimate a semiparametric hedonic housing price equation where the contribution of housing attributes to home prices is obtained parametrically, but the component of home prices that varies with location is not assumed to lie in a given parametric family.

Using data from 2002–2006, we apply this method to the city of Richmond, Virginia, and three surrounding counties. The region under consideration covers approximately 1,218 square miles, comprises nearly one million people, and has boundaries that are agricultural in nature. Since our technique uses home transaction sales, and not simply vacant land, we are able to construct a land gradient from over 100,000 observations. Surprisingly, given the recent trend toward polycentric cities in the United States, we find that the price surface in Richmond is largely monocentric, with land prices falling from over \$100 per square foot (in 2006 constant dollars) around the CBD to less than \$1 per square foot in the rural outskirts. Though the CBD is the dominant feature on the price surface, larger suburbs, such as Mechanicsville, Ashland, Short Pump, and Midlothian, and corridors along Interstates 64 and 95, are easily identified. Furthermore, the presence of these subcenters distorts estimates of parametric surfaces even when they assume one dominant center.

An exponential function fitted to our estimates of land prices reveals that prices fall, on average, at the rate of 2.8 percent per mile as one moves away from Richmond's CBD. Put another way, land prices fall by $\frac{1}{2}$ every 25 miles. This rate is significantly higher than the rate of decline estimated with a least squares regression of home prices on housing characteristics and a measure of distance from the CBD, which finds prices falling at only 1 percent per mile. This difference arises because conventional parametric methods do not allow for local variations in housing prices and, consequently, achieve a considerably worse fit over a large area. In particular, the parametric regression is associated with a much poorer fit of the data relative to its semiparametric counterpart.

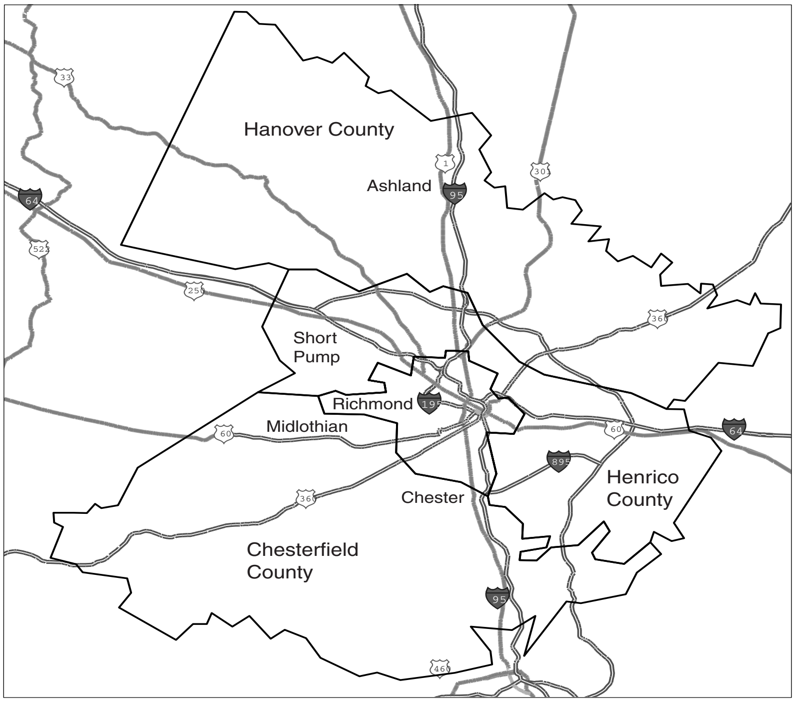
The technique for estimating land prices proposed in this article makes no assumptions about the geography of the Richmond area or the structure of the land price surface. In a parametric estimation of land prices, the inclusion of variables to represent location within a certain county or distance from an identified "employment subcenter" (such as Chicago O'Hare airport in Colwell and Munneke [1997]) is essential to achieving a reasonably accurate land price surface. However, identifying locations such as an employment subcenter can be a difficult and arbitrary task (Giuliano and Small [1991] and McMillen [2001]). Moreover, independent of commuting considerations, proximity to geographical features such as lakes may enhance the value of certain locations. By construction, this feature of land prices cannot be captured by parametric methods based on distance from a CBD. In contrast, because no arbitrary decisions about the functional form of the land surface need to be made with the method used in this article, it can be directly applied to any urban region of any shape and size.

The rest of the article proceeds as follows. In Section 1, we describe the data and features of the Richmond area. Section 2 describes the empirical model and discusses how semiparametric land price estimates are computed. In Section 3, we construct the land price surface and compare our estimates with those constructed using simpler polynomial and distance-from-CBD estimates. Section 4 concludes.

1. DATA DESCRIPTION

This article estimates the land price gradient from a full sample of residential sales in the city of Richmond and three nearby counties—Hanover, Henrico, and Chesterfield—from 2002–2006. Richmond is a mid-sized regional center with a population just over 200,000, lying 100 miles south of Washington, D.C., at the intersection of Interstates 95 and 64. The urban core was well developed by the late 19th century, when the city served as the Confederate capital during the Civil War. As such, the mean age of the housing stock in the city itself is more than 66 years. Hanover County lies due north of Richmond, with a largely rural population of less than 100,000, though significant suburbs do line the Interstate 95 corridor. Henrico County lies both to the west and to the east of Richmond, with a population just under 300,000, and is home to a number of quickly growing suburbs surrounding Interstate 64, notably the areas around Short Pump and Mechanicsville. Chesterfield County, with a population over 300,000, lies to the south of Richmond and is primarily made up of low density suburban areas, along with a few notable small towns such as Chester and Midlothian. A map of Richmond and surrounding counties is given in Figure 1. All told, the region includes almost one million people

Figure 1 Richmond and Surrounding Counties



residing in over 1,218.5 square miles. Aside from the far southern end of Chesterfield County, which abuts the cities of Colonial Heights and Petersburg, the edge of this region consists of rural farmland.

We acquired a full record of property sales, with matched housing characteristics, from the city and counties. These characteristics include the furnished square footage of a house, the number of years since the house was first built, its plot acreage, and the number of bathrooms available. We also include binary variables that indicate whether a house has air conditioning, whether its exterior is made of brick, vinyl, or wood, and whether it is heated using oil, hot water, or central air. Before carrying out the estimation, we filter the data along several dimensions. First, all nonresidential properties were removed, as the parametric portion of our estimation requires data on, for instance, livable square footage. Next, we remove houses that appear to have improperly entered data—this includes houses with construction dates before 1800, houses with sales prices of less than one dollar, houses that appear to have been sold in a lot where no breakdown of the sales price per house

Table 1 Data Summary

Variable	Mean	St. Dev.	Min.	Max.
Sales Price[a]	210,068.63	152,035.95	6.46	5,639,195.45
Age	33.14	28.61	1	207
No. of Bathrooms	2.13	0.90	0	72
Air Conditioning	0.66	0.47	0	1
Square Footage	1,876.0	903.4	319	63,233
Lot Acreage	0.70	2.44	0.02	98.77
Brick Exterior	0.20	0.40	0	1
Vinyl Exterior	0.31	0.46	0	1
Wood Exterior	0.18	0.38	0	1
Gas Heating	0.09	0.29	0	1
Oil Heating	0.17	0.38	0	1
Hot Water Heating	0.14	0.35	0	1
Central Air Heating	0.13	0.33	0	1

Notes: [a]Expressed in constant 2006 dollars.

is available, and houses with plot acreage lower than .02 acres (871 square feet).

We geocode data from street addresses in ArcGIS using SPCS NAD83 coordinates, which, unlike simple latitude and longitude, allow easy calculation of Euclidean distance in feet between any two points. In some cases, the geocoder was unable to positively identify a street address; unidentified houses are left out of the final sample. Descriptive statistics of our data are reported in Table 1.

2. THE EMPIRICAL MODEL

This section sets up the basic framework we use throughout the remainder of the article. We denote the city of Richmond (and its four surrounding counties) by C and a location in the city by $\ell = (x, y) \in C$, where x and y are Cartesian coordinates. We denote the (log) per-square-foot price of a home in Richmond by p. Our analysis begins with the following hedonic price equation,

$$p = \mathbf{X}\beta + f(\ell) + \varepsilon, \tag{1}$$

where $\mathbf{X}$ is a k-element vector of conditioning housing characteristics such that $cov(\mathbf{X}|\ell) = \Sigma_{\mathbf{X}|\ell}$, $f(\ell)$ is the component of a home price directly related to location, and ε is a random variable such that $E(\varepsilon|\ell, \mathbf{X}) = 0$ and $var(\varepsilon|\ell) = \sigma_{\varepsilon}^2$. The matrix $\mathbf{X}$ consists of all of the variables from Table 1 and quadratic terms for lot acreage and square footage, as Brownstone and De Vany (1991), among others, find that land price per acre is a concave function of parcel size. The coefficients, β, capture the reduced-form effects of particular housing attributes, such as the size of the living area of a house or whether air conditioning is available, on home prices. Moreover, since $p - \mathbf{X}\beta$ represents housing prices purged of the contribution from specific attributes, we think of $f(\ell)$ as capturing the value of land per-square-foot at a given location. While this general semi-log specification is standard in the analysis of real estate data, some differences exist regarding the functional form that describes the function $f(.)$. One option is to specify $f(.)$ as a polynomial function of location data, as in Galster, Tatian, and Accordino (2006),

$$f(x, y) = \alpha_0 x + \alpha_1 y + \alpha_3 x^2 + \alpha_4 y^2 + \alpha_5 xy. \tag{2}$$

Substituting equation (2) into equation (1), one can then consistently estimate the coefficients β and α_i, $i = 1, .., 5$, using least squares.

An alternative approach that uses least squares estimation is to parameterize $f(.)$ as a function of distance from the CBD, as in Zheng and Khan (2008),

$$f(x, y) = \alpha_d \sqrt{(x - x_c)^2 + (y - y_c)^2}, \tag{3}$$

where $\ell_c = (x_c, y_c)$ denotes the location of the CBD. Recalling that p is measured in log units, α_d then captures the exponential rate of change in land values as one moves away from the CBD.

In contrast to either of these approaches, this article does not assume that $f(\ell)$ lies in a given parametric family. The only restriction that we shall impose on $f(\ell)$ is that it satisfies a Lipschitz condition,

$$||f(\ell_a) - f(\ell_b)|| < L||\ell_a - \ell_b||, L \geq 0. \tag{4}$$

Semiparametric Regression

Estimating the nonparametric component of equation (1), $f(\ell)$, requires that we first address estimation of the parametric effects, β. One strategy would be to estimate equation (1) in two stages, first ignoring the nonparametric component, $f(\ell)$, in order to obtain estimates of β by regressing p on $\mathbf{X}$, and then applying nonparametric methods to purged home prices, $p - \mathbf{X}\hat{\beta}$, where $\hat{\beta}$ denotes the previously obtained estimates of β. However, since the reduced form model contains a component related to location that is being ignored, estimates of β obtained in this way will be biased when housing attributes, $\mathbf{X}$, are correlated with location, ℓ. Rather, a two-step estimation strategy must somehow "get rid" of the nonparametric component in the first step.

Let n denote the number of observations in our data set. A popular approach, pioneered by Robinson (1988), recognizes as a first step that equation (1) implies that

$$p - E(p|\ell) = [\mathbf{X} - E(\mathbf{X}|\ell)]\beta + \varepsilon. \tag{5}$$

In other words, the conditional differencing of equation (1) gets rid of the nonparametric component. Robinson (1988) then shows that by replacing $E(p|\ell)$ and $E(\mathbf{X}|\ell)$ with nonparametric kernel estimates (to be described below) $\widehat{E}(p|\ell)$ and $\widehat{E}(\mathbf{X}|\ell)$, respectively, and then regressing $p- \widehat{E}(p|\ell)$ on $[\mathbf{X} - \widehat{E}(\mathbf{X}|\ell)]$, yields estimates of β that are $\sqrt{n}$ consistent. Unfortunately, this method can be quite onerous since separate nonparametric regressions are required for each housing attribute in $\mathbf{X}$, where both the number of relevant housing attributes and the number of observations are large in our case. To avoid this problem, we summarize instead a differencing method developed more recently by Yatchew (1997) and Yatchew and No (2001), and adopted, for example, in Rossi-Hansberg, Sarte, and Owens (2008).

The basic idea behind differencing the data works as follows. We would like to re-order our data, $(p_1, \mathbf{X}_1, \ell_1), (p_2, \mathbf{X}_2, \ell_2), ... , (p_n, \mathbf{X}_n, \ell_n)$ so that the ℓ's are close, in which case differencing tends to remove the nonparametric effects. To get a sense of the implications of differencing, suppose that locations constitute a uniform grid on the unit square (the re-scaling is without loss of generality). Each point may then be thought of as taking up an area of $\frac{1}{n}$, and the distance between adjacent observations is therefore $\frac{1}{\sqrt{n}}$. Suppose further that the data have been re-ordered so that $||\ell_i - \ell_{i-1}|| = \frac{1}{\sqrt{n}}$. First-differencing of (1) then yields

$$p_i - p_{i-1} = (\mathbf{X}_i - \mathbf{X}_{i-1})\beta + f(\ell_i) - f(\ell_{i-1}) + \varepsilon_i - \varepsilon_{i-1}. \tag{6}$$

Assuming that equation (4) holds, the difference in nonparametric components in (6) vanishes asymptotically. Yatchew (1997) then shows that the ordinary least squares estimator of β using the differenced data (i.e., by regressing $p_i - p_{i-1}$ on $\mathbf{X}_i - \mathbf{X}_{i-1}$) is also $\sqrt{n}$ consistent. This estimator of β, however, achieves only $\frac{2}{3}$ efficiency relative to the one produced by Robinson's method. This can be improved dramatically by way of higher-order differencing. Specifically, define $\Delta\mathbf{p}$ to be the $(n-m) \times 1$ vector whose elements are $[\Delta\mathbf{p}]_i = \sum_{s=0}^{m} \omega_s p_{i-s}$, $\Delta\mathbf{X}$ to be the $(n-m) \times k$ matrix with entries $[\Delta\mathbf{X}]_{ij} = \sum_{s=0}^{m} \omega_s X_{i-s,j}$, and similarly for $\Delta\varepsilon$. The parameter m governs the order of differencing and the ω's denote constant differencing weights. Equation (6) can then be generalized as

$$\Delta\mathbf{p} = \Delta\mathbf{X}\beta + \sum_{s=0}^{m} \omega_s f(\ell_{i-s}) + \Delta\varepsilon, \ i = m+1, ..., n, \tag{7}$$

where the following two conditions are imposed on the differencing coefficients, $\omega_0, ..., \omega_m$:

$$\sum_{s=0}^{m} \omega_s = 0 \text{ and } \sum_{s=0}^{m} \omega_s^2 = 1. \tag{8}$$

The first condition ensures that differencing removes the nonparametric effect in (1) as the sample size increases and the re-ordered ℓ's get closer to each other. The second condition is a normalization restriction that implies that the variance of the transformed residuals in (7) is the same as the variance of the residuals in (1). When the differencing weights are chosen optimally, the difference estimator β_Δ obtained by regressing $\Delta\mathbf{p}$ on $\Delta\mathbf{X}$ approaches asymptotic efficiency by selecting m sufficiently large.[1] In particular, Yatchew (1997) shows that

$$\hat{\beta}_\Delta \sim^A N(\beta, (1+\frac{1}{2m})\frac{\sigma_\varepsilon^2}{n}\Sigma_{\mathbf{X}|\ell}^{-1}),$$

$$s_\Delta^2 = \frac{1}{n}\sum_{i=1}^{n}(\Delta\mathbf{p}_i - \Delta\mathbf{X}_i\hat{\beta})^2 \to^P \sigma_\varepsilon^2, \text{ and} \qquad (9)$$

$$\widehat{\Sigma}_{u,\Delta} = \frac{1}{n}\Delta\mathbf{X}'\Delta\mathbf{X} \to^P \Sigma_{\mathbf{X}|\ell}.$$

These results will allow us to do inference on $\hat{\beta}_\Delta$. By equation (9), the R^2 statistic associated with our original empirical specification (1) can be estimated as $R^2 = 1 - \frac{s_\Delta^2}{s_p^2}$. In our estimation exercise, we use $m = 10$, which produces coefficient estimates that are approximately 95 percent efficient when using optimal differencing weights.

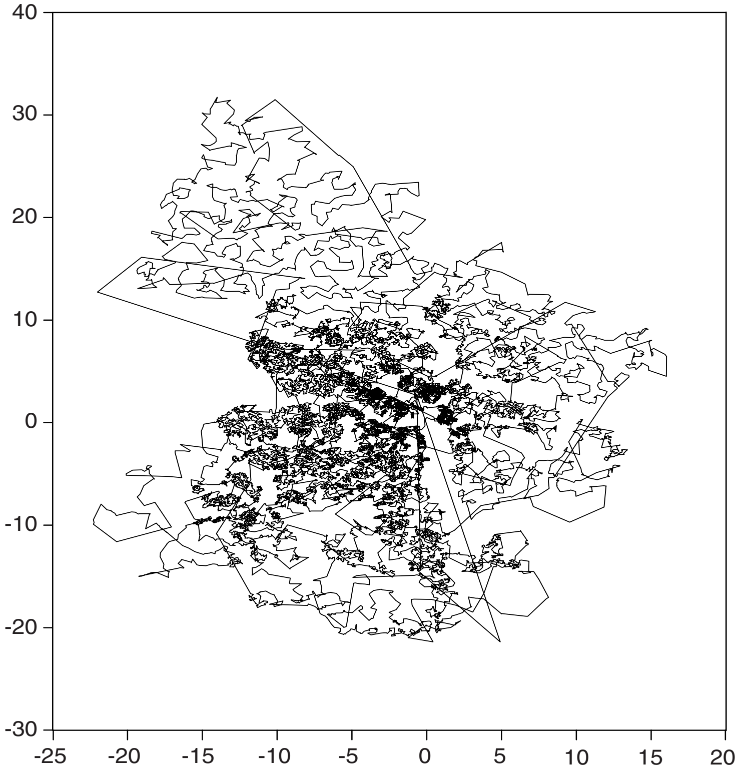
Finally, note that because locations lie in R^2, the initial re-ordering of the ℓ's is not unambiguous in this case. A Hamiltonian cycle over distance is the ordering of housing sale coordinates such that the sum of differenced distances is minimized. However, computing a Hamiltonian cycle for over 100,000 points is not yet tractable on a personal computer.[2] In this article, we re-order locations using a nearest-neighbor algorithm that finds an approximate Hamiltonian cycle in the following way: First, we select an arbitrary starting location from which we then find the location of a sale in our data set nearest to it in Euclidean distance. From this second location, we then find the nearest third sale location among the set of remaining observations (i.e., those not already identified). This process is repeated until every sale location has been selected.[3] Figure 2 displays the path chosen by our algorithm. The median distance between points in our sample is 86 feet. As this figure is darker in areas where there are more residential sales, it also serves as a rough guide to residential density in the Richmond area. The origin represents the CBD of Richmond City.

[1] Optimal differencing weights, $\omega_0, ..., \omega_m$, solve min $\delta = \sum_{k=1}^{m}(\sum_s \omega_s\omega_{s+k})^2$ subject to the constraints in (7). See Proposition 1 in Yatchew (1997).

[2] The Hamiltonian cycle is the solution to the famous Traveling Salesman Problem (TSP). As of 2007, the largest TSP ever solved on a supercomputer involved 85,900 points, which is smaller than our problem (Applegate et al. 2006).

[3] See Rosenkrantz, Stearns, and Lewis (1977). Although the starting point is arbitrary, it has little implications for our findings.

Figure 2 Path of Approximate Hamiltonian Cycle



Nonparametric Kernel Estimation of $f(\ell)$

Denote by z the price of a home "purged" of its contribution from housing characteristics, where z is obtained using first stage estimates, $z = p - \mathbf{X}\hat{\beta}_{\Delta}$, and construct the data (z_1, ℓ_1), (z_2, ℓ_2),...,(z_n, ℓ_n). Then, because $\hat{\beta}_{\Delta}$ is a consistent estimator of β, the consistency of $f(\ell)$ obtained using standard kernel estimation methods applied to purged home prices remains valid.

The *Nadaraya-Watson* kernel estimator of f at location ℓ_j is given by

$$f(\ell_j) = n^{-1} \sum_{i=1}^{n} W_{hi}(\ell_j) z_i. \tag{10}$$

In other words, the value of land at location ℓ_j is a weighted-average of the z's in our data sample. The weight, $W_{hi}(\ell_j)$, attached to each purged price,

z_i, is given by

$$W_{hi}(\ell_j) = \frac{K_h(\ell_j - \ell_i)}{n^{-1}\sum_{i=1}^{n} K_h(\ell_j - \ell_i)}, \tag{11}$$

where

$$K_h(u) = h^{-1}K(\frac{u}{h}),$$

and $K(\psi)$ is a symmetric real function such that $\int |K(\psi)|d\psi < \infty$ and $\int K(\psi)d\psi = 1$. Thus, we may choose to attach greater weight to observations on prices of homes located near ℓ_j rather than far away by suitable choice of the function K. In particular, as in much of the literature, our estimation is carried out using the Epanechnikov kernel,

$$K(\frac{u}{h}) = \frac{3}{4}\left(1 - \left(\frac{u}{h}\right)^2\right) I(\left|\frac{u}{h}\right| \leq 1), \tag{12}$$

where $I(.)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The distance between location ℓ_j and some other location ℓ_i in Richmond is simply measured as a Euclidean distance in feet. The kernel in (12) then implies that prices of homes located more than a distance of h feet from ℓ_j will receive a zero weight in the estimation of $f(\ell_j)$. In that sense, the bandwidth, h, has a very natural interpretation in this case. In practice, the estimation of $f(\ell)$ is affected to a greater degree by the choice of bandwidth rather than the choice of kernel.[4]

How then does one choose what bandwidth is appropriate? A seemingly natural method for choosing the bandwidth is to minimize the sum of squared residuals,

$$MSE = n^{-1}\sum_{i=1}^{n}[z_i - f(\ell_i)]^2.$$

However, because z_i is used when estimating $f(\ell_i)$, the mean squared error can be arbitrarily reduced by decreasing the bandwidth until all weight in $f(\ell_i)$ is effectively placed on z_i. To avoid this problem, the cross-validation method proposes that the bandwidth parameter be chosen by minimizing the sum of squared residuals from an alternative kernel regression in which z_i is dropped in the estimation of $f(\ell_i)$. Hence, we select h so that it solves

$$\min_h CV(h) = n^{-1}\sum_{i=1}^{n}[z_i - \widetilde{f}_h(\ell_i)]^2, \tag{13}$$

[4] See DiNardo and Tobias (2001).

where

$$\widetilde{f}_h(\ell_j) = n^{-1} \sum_{i \neq j}^{n} W_{hi}(\ell_j) z_i .$$

We estimate equation (7) using 103,543 observations over the period 2002–2006. All prices are deflated using the consumer price index and measured in 2006 constant dollars. We include among our conditioning variables, **X**, a set of time dummies associated with the sale date of a home that captures secular citywide increases in real home prices, where 2006 is set as the base year.

3. FINDINGS

This section reviews our findings. Before describing the results from the semiparametric estimation, we first present estimates from the polynomial specification of Galster, Tatian, and Accordino (2006), equation (2), and the parameterized distance function of Zheng and Khan (2008), equation (3), as benchmarks.

Table 2 presents estimates from the specification where the value of land is modeled as a quadratic function of location data, as in equation (2), under the heading "Parametric Model 1." The estimation of the coefficients is carried out using least squares. Virtually all housing characteristics are statistically significant at the 5 percent critical level and most are significant at the 1 percent level. The coefficients associated with the sale date are significant over and above prices being measured in constant dollars. In particular, the findings suggest a general increase in real home prices over our sample period (recall that 2006 is set as the base year). In addition, the regression achieves a relatively good fit for cross-sectional data, with an R^2 of about .50.

Of central interest in the first two columns of Table 2 are the parameters that govern the value of land associated with the Cartesian location data. The coefficients of the polynomial in equation (2) are all highly significant, with the exception of the cross-term, which is statistically significant at the 10 percent critical level only. Figure 3 shows the land value surface associated with these parameters. The origin roughly represents the CBD of the city of Richmond, at the intersection of 7th Street and Canal Street. This location corresponds to coordinates within an area generally considered the employment center of Richmond, with high employment density and a preponderance of commercial and office buildings. The polynomial estimate of land prices, however, has a peak nearly 20 miles away from the CBD, roughly located in a far western suburb known as Short Pump. Since the polynomial in equation (2) permits, at most, one local interior maximum, this parametric regression imposes a land price surface typical of a monocentric city. Essentially, the polynomial will choose a maximum in an area for which there exist many house sales with a

Table 2 Modeling Land Prices as Functions of Local Data

	Parametric Model 1		Parametric Model 2		Semiparametric Model	
Variable	**Coeff.**	***t*-Statistics**	**Coeff.**	***t*-Statistics**	**Coeff.**	***t*-Statistics**
2002	−0.41	−52.37	−0.40	−50.10	−0.41	−74.64
2003	−0.32	−42.09	−0.31	−40.11	−0.33	−60.93
2004	−0.22	−29.32	−0.22	−27.81	−0.23	−43.07
2005	−0.09	−12.39	−0.09	−11.37	−0.10	−19.13
Age[a]	0.30	25.40	0.26	20.37	−0.40	−29.65
No. of Bathrooms	0.13	30.81	0.16	35.21	0.04	11.76
Air Conditioning	0.33	61.35	0.36	65.07	0.10	21.26
Sq. Ft.[b]	0.15	31.55	0.18	35.55	0.11	25.29
(Sq. Ft.)^{2}[c]	−0.38	−20.37	−0.45	−23.27	−0.25	−17.61
Acreage	−0.39	−222.58	−0.40	−225.31	−0.31	−161.19
(Acreage)^{2}[d]	0.54	123.50	0.56	124.35	0.38	106.39
Brick Exterior	0.04	6.02	0.04	5.52	−0.01	−1.79
Vinyl Exterior	0.06	10.05	−0.01	−1.48	−0.02	−4.37
Wood Exterior	−0.08	−11.01	−0.07	−9.21	−0.02	−3.99
Gas Heating	0.22	25.16	0.21	22.94	0.09	12.79
Oil Heating	0.04	4.56	0.18	25.03	0.06	6.96
Hot Water Heating	0.24	29.21	0.36	46.12	0.07	8.25
Central Air Heating	0.17	21.67	0.20	23.34	0.03	4.06
x	2.45	10.04				
y	−3.65	−31.88				
x^2	−0.09	−13.87				
y^2	−0.13	−37.07				
xy	−0.03	−3.88				
Distance from CBD			−0.01	−14.63		
R^2	0.50		0.47		0.77	

Notes: [a]Measured in 100 years; [b]measured in 1,000 sq. ft.; [c]measured in 100 million sq. ft. squared; [d]measured in 100 acres squared.

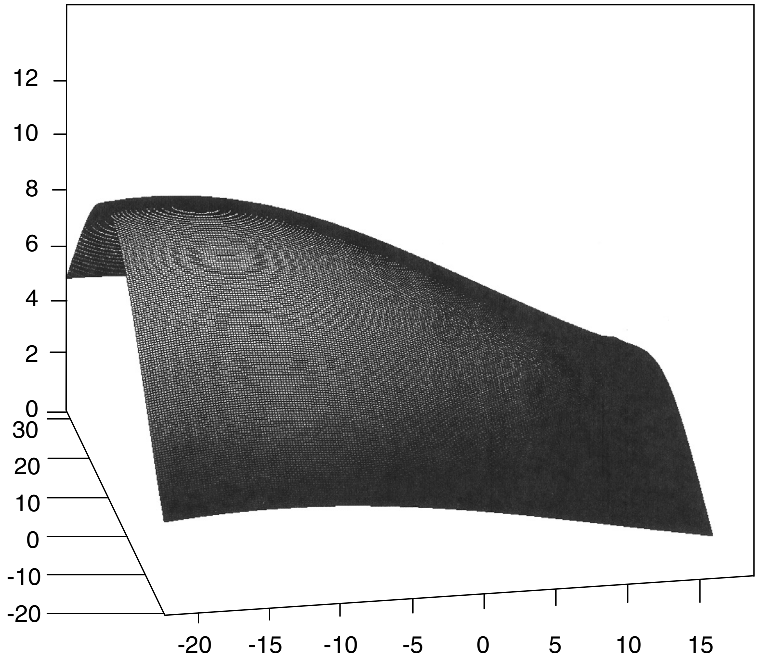
fairly high land value, even if there are other local maxima (near the CBD) on the true land price surface with higher prices but fewer sales.

Table 2 also presents estimates from the parametric specification where land prices (per square foot) are assumed to decline at an exponential rate with distance from the CBD, equation (3), under the heading "Parametric Model 2." The results generally mimic those of our first parametric model, both in terms of the coefficients associated with the different housing attributes and their statistical significance. For example, an additional bathroom adds approximately \$0.13 to the log per-square-foot price of a home under the first specification in Table 2 and \$0.16 under the second parameterization. This translates to an additional bathroom, adding \$2.39 and \$2.68, respectively, to the per-square-foot price of an average home. Note, however, that the parametric specification for Model 2 achieves a slightly worse fit than that for Model 1, with an R^2 statistic of 0.47.

Under the parametric specification including distance from the CBD, we find that the log price per square foot of land declines at a rate of .0103 per mile as one moves away from Richmond's CBD. This translates into a price per square foot of land that falls exponentially at a rate of 1.03 percent per mile as distance from the CBD increases. Alternatively, this exponential rate of decay implies that land prices fall by $\frac{1}{2}$ approximately every 67 miles. This rate of decline in land prices is significantly slower than those estimated by Zheng and Khan (2008) for Beijing, and Colwell and Munneke (1997) for Chicago. The difference occurs because our estimation exercise extends over an area much larger than that covered in the latter two papers, extending 50 miles in diameter in our case, using a similar specification. Because differences in geography are potentially much more pronounced over a larger area, the restriction embedded in the parametric specification related to location will be more stringent and its fit becomes poorer. Figure 4 shows the land price surface associated with our second parametric model. By construction, given the specification in equation (3), this surface reaches a peak at the location that we defined as the center of the CBD, $\ell_c = (x_c, y_c)$ in equation (3). As in the other parametric regression estimated above, this specification imposes a single peak on the land price surface as expected for a monocentric city. Compared to Figure 3, Figure 4 suggests considerably less variation in land prices throughout the city, with a peak price of \$7.60 per square foot at the CBD and approximately \$3.80 at the boundaries of greater Richmond. These figures translate to \$33,106 and \$16,553, respectively, for a 0.1 acre lot.

Given the variation in home prices in greater Richmond depicted in Table 1, this relatively narrow range in estimated land prices implied by Model 2 is potentially surprising. Moreover, the fact that these estimates stem from a parametric regression whose fit is slightly worse than that associated with the first parametric model in Table 2 should leave us somewhat skeptical. As we now discuss, the alternative specification where $f(\ell)$ in equation (1) is treated

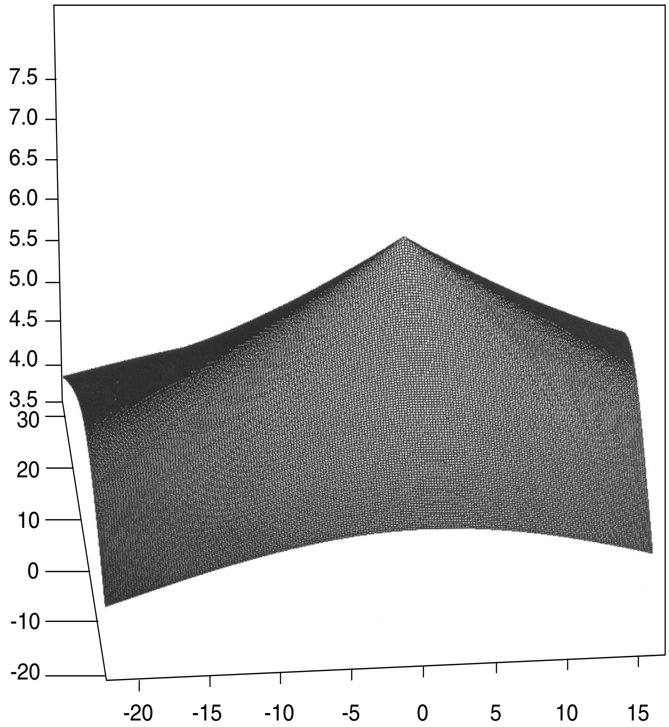
Figure 3 Land Price Surface Estimated from Parametric Model I



nonparametrically yields a significantly better fit, and implies a much more varied and greater range in land prices.

In contrast to the findings from the parametric approaches we have just reviewed, the last two columns of Table 2 present estimates from the semiparametric method described in Section 2. As before, virtually all variables are statistically significant at the 1 percent critical level, but the magnitude of the coefficients differs somewhat from those of our first two parametric specifications. For example, an additional bathroom now contributes .038 to the log per-square-foot price of a home as opposed to .13 for Model 1 and .16 for Model 2. The difference stems from the fact that we now estimate the component of home prices associated with location nonparametrically. In particular, observe that this semiparametric specification achieves a noticeably better fit relative to the previous two parametric specifications with an R^2 statistic of 0.77 instead of 0.50. Put another way, the semiparametric method

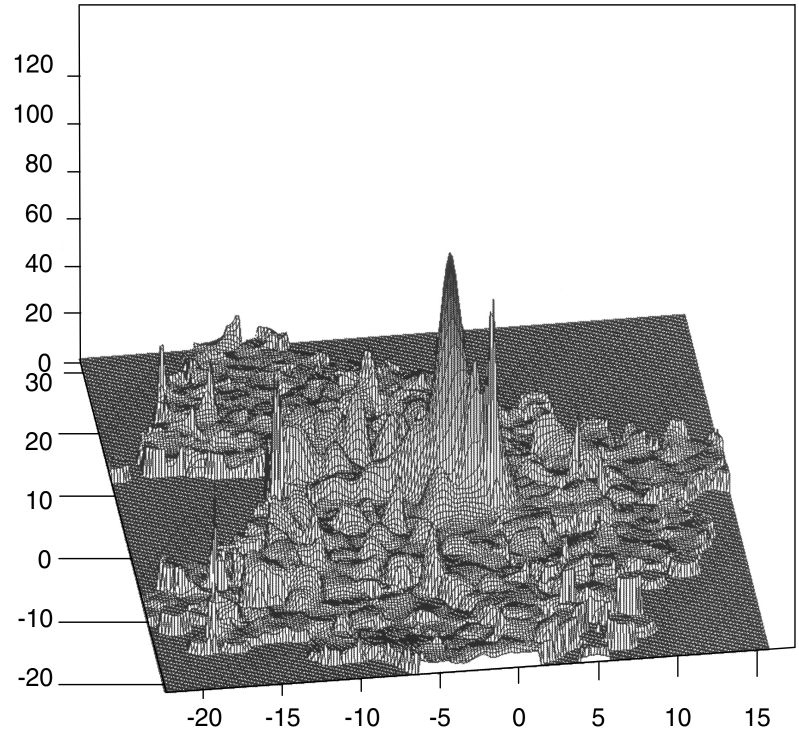
Figure 4 Land Price Surface Estimated from Parametric Model II



adopted here improves the fit of the parametric regressions carried out earlier by almost 60 percent.

Figure 5 illustrates the land price surface obtained using kernel estimation. Evidently, this surface differs considerably from those shown in Figures 3 and 4 along at least two important dimensions. First, this surface displays more variation in land prices across different areas of Richmond than could be obtained from parametric estimates. Observe, for instance, that the West End of Richmond is generally characterized by higher land prices than the area east of the city. The surface also displays multiple local peaks in prices associated with different parts of the city. Second, although the nonparametric estimation identifies one main peak, land prices where this peak is located are as high as $130 per square foot, in contrast to $10 per square foot using the polynomial specification in (2). A typical 0.1 acre lot in the most expensive neighborhoods of Richmond, therefore, is estimated at around $566,280 as opposed to $43,560 obtained earlier with the polynomial parameterization.

Figure 5 Land Price Surface Estimated from Semiparametric Model

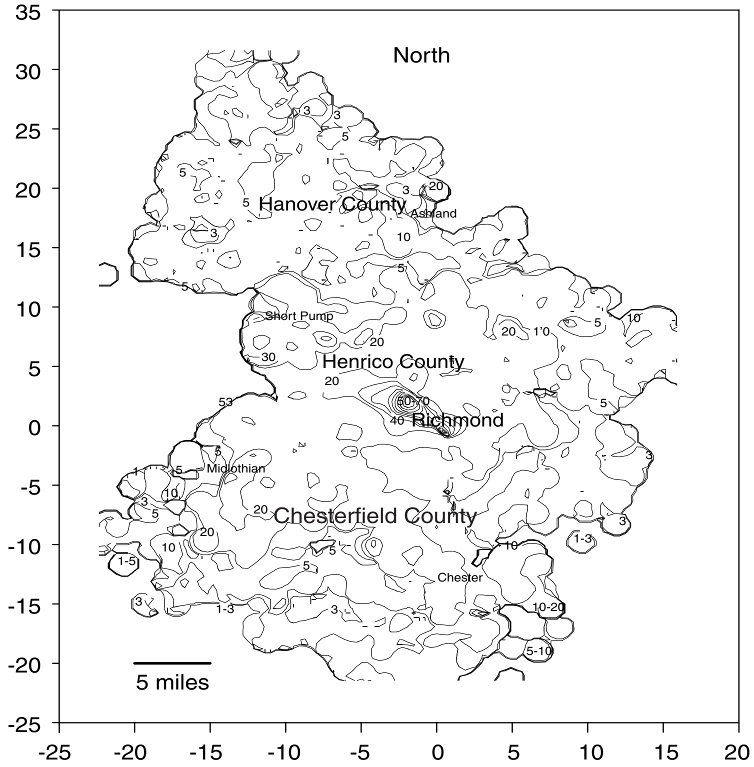


The main reason underlying these differences arises because nonparametric estimation relies on local averaging of the data—sharp peaks and valleys are much more easily discovered with a nonparametric estimation. More specifically, the bandwidth that minimizes the cross-validation criterion in equation (13) is around 5,000 feet in our case. In other words, in estimating land prices at any given location, our procedure uses data within 5,000 feet of that location, with weights that decay quadratically in (12) as one moves away from the point of estimation.

Figure 6 shows the contour map corresponding to the land price surface shown in Figure 5. A main peak is clearly visible just to the north and west of the CBD and corresponds to an area of expensive row houses known as the Fan District in Richmond. Prices in that neighborhood are as high as $80 to $130 per square foot.

In contrast, land prices near the boundaries of Richmond range from only $1 to $2 per square foot and capture the opportunity cost of land related to

Figure 6 Contour Map of Land Price Surface



agricultural activity at those locations. Local peaks in prices in Figure 6 are also visible five to 15 miles north and west of the city, around areas known as Short Pump and the West End more generally. These areas lie around Interstate 64 and consist of a number of newer suburban neighborhoods generally made up of single-family homes. The contour plot also shows evidence of local peaks extending north from the CBD around the Interstate 95 corridor and mid-sized towns such as Ashland and Mechanicsville. Finally, a series of local maxima can be found 12 miles west and 6 miles south of the CBD, in a region featuring a number of small lakes and golf courses.

Interestingly, despite the variations in land values shown in Figures 5 and 6, our findings suggest that Richmond remains largely a monocentric city. The more recent expansion in activity in the areas of Short Pump, west of the city, is associated with higher land prices on average compared to other areas located a similar distance away from Richmond's CBD. On the whole,

Table 3 Estimates versus Assessments

(Per Square Foot)	Assessments	Estimate
Mean	2.46	9.50
Median	2.02	9.15
Maximum	30.75	33.89
Standard Deviation	1.98	5.18

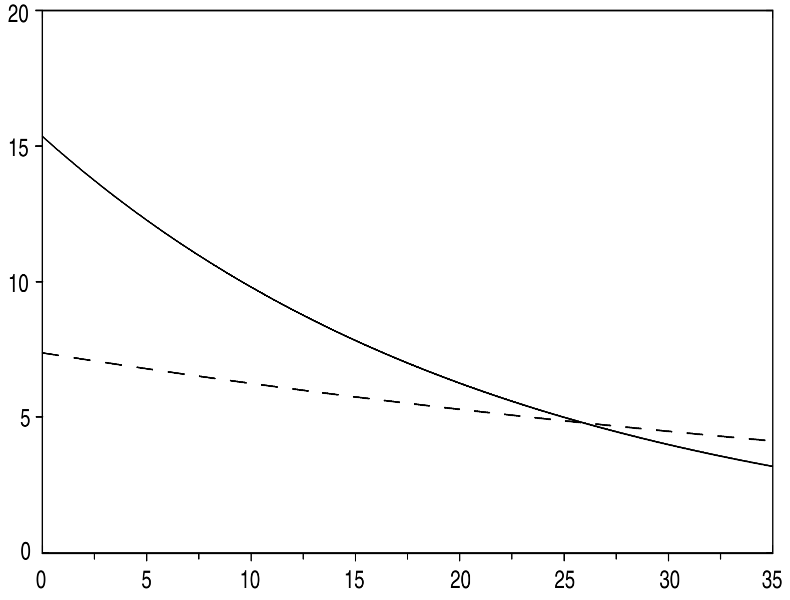
however, land values are highest in the older sections of Richmond near the center of the city.

To compare our estimated land prices with alternative estimates, we obtain land price assessments from Chesterfield County, the county that lies in the southern portion of our region.[5] For tax purposes, Chesterfield County computes assessments both for underlying land value and for improvements. The land value assessment is updated every year and is based on "comparable" vacant land sales from other regions in the county. Though this method allows for much less local variation than our semiparametric approach, it provides a rough estimate of land costs per square foot in the county. Over the period studied, land assessments average just under $3 per square foot, with a number of plots assessed at under $1 per square foot, and the most expensive county land assessed at $20 to $30 per square foot. Table 3 displays characteristics of each distribution. The most expensive plots were located near the golf courses and lakes southwest of Richmond's CBD that were identified as a local maximum in the semiparametric estimation. The assessed value of land tends to be $3 to $6 per square foot less than our semiparametric estimation. This difference reflects, in part, housing characteristics that we are unable to control for in the first step of the nonparametric estimation. For data availability reasons, we cannot remove every piece of a house—for instance, there is no dummy for the number of fireplaces, whether the lot is fenced, the particular layout of the house, the quality of interior materials, etc. The fact that a house exists on every land plot in our sample means that our definition of land is more precisely that of a plot with a zero square foot, zero bathroom house with some unidentified exterior and heating type. Practically, this means that houses in our sample have already installed utility hookups and already have the appropriate zoning for a residential house.

Unless the house is of very poor quality, a plot with those characteristics will be more valuable than empty land, and, therefore, our nonparametric land price estimates will be somewhat biased upwards. However, whatever bias

[5] In many counties, full records of assessment data are not free to obtain; the Chesterfield Assessment Office, however, was able to send us land assessments linked to our housing sales data.

Figure 7 Estimated Land Price Gradient



exists will be evident across all locations, so that differences in the price of land at one location relative to other locations should be similar in both the land assessments and our estimated land price; this indeed appears to be the case for Chesterfield County.

Finally, recall that the parametric specification, including distance from the CBD, delivers a small range of land values and a very shallow price gradient. Figure 7 shows an estimate of the land gradient obtained by projecting our estimates of (log) per-square-foot land prices at a given location onto distance from the CBD at that location. Put another way, we estimate the following equation,

$$\widehat{f}(\ell) = \alpha_0 + \alpha_d d(\ell) + e, \tag{14}$$

where $\widehat{f}(\ell)$ denotes our nonparametric estimate of (log) land value at location ℓ, $d(\ell)$ is the distance to the center of the CBD from ℓ, and α_d represents an exponential rate of change in prices. Our findings suggest a substantially faster rate of decay in land prices (the solid line) than estimated previously for Richmond using the parametric specification in equation (3) (the dashed line). The parametric specification estimated earlier gave a decline of 1.03

percent per mile. In contrast, we now find that α_d is approximately $-.0278$, as opposed to $-.0103$ in Table 2. This implies that land prices decline at a rate of 2.78 percent per mile on average as one moves away from Richmond's CBD. Alternatively, land prices fall by $\frac{1}{2}$ approximately every 25 miles as distance from the CBD increases.

4. CONCLUDING REMARKS

The complexity of the urban price surface means that the assumptions that prices decline monotonically from a CBD or reflect a simple polynomial function of location data are not innocuous. Transit corridors, bodies of water, parkland, golf courses, employment subcenters, and other topographical features can have significant effects on land prices around a city. While these features could in theory be controlled for, it is not straightforward to identify what features or employment centers might be worth identifying.

Drawing on recent work by Yatchew (1997) and Yatchew and No (2001), the semiparametric procedure outlined in this article allows for an approach that does not require a priori assumptions regarding what features of the landscape might affect land prices. It also allows a very large data set—that of all housing transactions in a region—to be used when estimating the land price gradient. Since this procedure does not, unlike earlier work on land prices, rely on local knowledge, it can be applied wholesale to any region or city.

Empirically, an application of this semiparametric approach to land price estimation in Richmond, Virginia, identifies local maxima in the land price surface principally along the Interstate 64 and 95 corridors, in the suburbs of Ashland and Short Pump, and around the lakes and golf courses south of Midlothian. The most expensive land in the region, by a large margin, lies in the historic district of the Fan located close to the CBD; prices in the Fan per acre are over 100 times more expensive than rural land in the surrounding counties.

REFERENCES

Anas, Alex, Richard Arnott, and Kenneth A. Small. 1998. "Urban Spatial Structure." *Journal of Economic Literature* 36 (September): 1,426–64.

Applegate, David L., Robert E. Bixby, Vasek Chvátal, and William J. Cook. 2006. *The Traveling Salesman Problem: A Computational Study*. Princeton, N.J.: Princeton University Press.

Brownstone, David F., and Arthur De Vany. 1991. "Zoning, Returns to Scale, and the Value of Undeveloped Land." *Review of Economics & Statistics* 73 (November): 699–704.

Colwell, Peter F., and Henry J. Munneke. 1997. "The Structure of Urban Land Prices." *Journal of Urban Economics* 41 (May): 321–36.

Colwell, Peter F., and Henry J. Munneke. 2003. "Estimating a Price Surface for Vacant Land in an Urban Area." *Land Economics* 79 (February): 15–28.

DiNardo, John, and Justin Tobias. 2001. "Nonparametric Density and Regression Estimation." *Journal of Economic Perspectives* 15 (Fall): 11–28.

Fujita, Masahisa. 1989. *Urban Economic Theory*. Cambridge, U.K.: Cambridge University Press.

Galster, George, Peter Tatian, and John Accordino. 2006. "Targeting Investments for Neighborhood Revitalization." *Journal of the American Planning Association* 72 (Autumn): 457–74.

Giuliano, Genevieve, and Kenneth A. Small. 1991. "Subcenters in the Los Angeles Region." *Regional Science and Urban Economics* 21 (July): 163–82.

McMillen, Daniel P. 2001. "Nonparametric Employment Subcenter Identification." *Journal of Urban Economics* 50 (November): 448–73.

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

Redfearn, Christian. 2007. "The Topography of Metropolitan Employment: Identifying Centers of Employment in a Polycentric Urban Area." *Journal of Urban Economics* 61 (May): 519–41.

Robinson, P.M. 1988. "Root-N-Consistent Semiparametric Estimation." *Econometrica* 56: 931–54.

Rosenkrantz, Daniel J., Richard E. Stearns, and Philip M. Lewis II. 1977. "An Analysis of Several Heuristics for the Traveling Salesman Problem." *SIAM Journal of Computing* 6: 563–81.

Rossi-Hansberg, Esteban, Pierre-Daniel G. Sarte, and Raymond E. Owens, III. 2008. "Housing Externalities: Evidence from Spatially Concentrated Urban Revitalization Programs." Federal Reserve Bank of Richmond Working Paper 08-03.

Seyfried, Warren R. 1963. "The Centrality of Urban Land Values." *Land Economics* 39 (August): 275–84.

Yatchew, Adonis. 1997. "An Elementary Estimator of the Partial Linear Model." *Economics Letters* 57 (December): 135–43.

Yatchew, Adonis, and Joungyeo Angelo No. 2001. “Household Gasoline Demand in Canada.” *Econometrica* 69 (November): 1,697–709.

Zheng, Siqi, and Matthew E. Khan. 2008. “Land and Residential Property Markets in a Booming Economy: New Evidence from Beijing.” *Journal of Urban Economics* 63 (March): 743–57.