HEDONIC PRICE FUNCTIONS: GUIDANCE ON EMPIRICAL SPECIFICATION

Nicolai V. Kuminoff
Applied Economics
Virginia Tech
540.231.5382
kuminoff@vt.edu

Christopher F. Parmeter
Applied Economics
Virginia Tech
540.231.0770
parms@vt.edu

Jaren C. Pope*
Applied Economics
Virginia Tech
540.231.4730
jcpope@vt.edu

Selected Paper Prepared for Presentation at the American Agricultural Economics Association Annual Meeting: Orlando, Florida, July 27-29, 2008

* Corresponding author.

HEDONIC PRICE FUNCTIONS: GUIDANCE ON EMPIRICAL SPECIFICATION

ABSTRACT: The hedonic pricing model is widely accepted as a method for estimating the marginal willingness to pay for spatially delineated amenities. Empirical applications typically rely on one of three functional forms—linear, semi-log, and double-log—and rarely involve rigorous specification testing. This phenomenon is largely due to an influential simulation study by Cropper, Deck and McConnell (CDM) (1988) that found, among other things, that simpler linear specifications outperformed more flexible functional forms in the face of omitted variables. In the 20 years that have elapsed since their study, there have been major computational advances and significant changes in the way hedonic price functions can be estimated. The purpose of our paper is to update and extend the CDM (1988) simulations to investigate current issues in hedonic modeling. Three preliminary results obtained from our theoretically consistent Monte Carlo simulation have been highlighted in this paper: (i) we find that adding spatial fixed effects (census tract dummies) to linear models does improve their performance. This is true both when all attributes are observed, and when some attributes are unobserved, (ii) adding the spatial fixed effects to the more flexible specifications such as the quadratic and quadratic box-cox does not improve their performance when all housing attributes are observed. However, when some housing attributes are unobserved, the spatial fixed effects significantly improves the performance of flexible specifications as well, and (iii) increasing the sample size from CDM's 200 observations to a sample size of 2000 (which is more representative of modern applications) changes the relative performance of different specifications.

KEY WORDS: Hedonic, Functional Form, Monte Carlo Simulation, Property Value Model

JEL CODES: Q15, Q51, Q53, C15, R52

1. Introduction

The hedonic pricing model is widely accepted as a method for estimating the marginal willingness to pay for spatially delineated amenities. It has been described by Palmquist and Smith (2002) as "one of the 'success' stories of modern applied micro-economic analysis." The method is frequently used to investigate important questions in agricultural, environmental, and urban economics. For example, hedonic pricing models have been used by Schlenker et al. (2005) to estimate the impact of climate change on farmland values, by Palmquist and Danielson (1989) to understand the value of erosion control and drainage, and by Smith and Huang (1995) to estimate the willingness to pay for marginal changes in air quality. Furthermore, recent quasi-experimental papers highlight the expanding role of the property value hedonic in evaluating public policies and the marginal-willingness-to-pay for public goods and spatially delineated amenities (i.e. Chay and Greenstone (2005), Davis (2004), and Pope (2007)). We expect the hedonic method will continue to play a prominent role in future empirical applications aimed at revealing household preferences for spatially delineated attributes that are tied to property markets.

From an empirical perspective, a key limitation of the hedonic method is that the underlying theory provides relatively little guidance on the shape of the equilibrium hedonic price function. Therefore, our ability to identify consumers' marginal willingness-to-pay (MWTP) for an amenity hinges on our maintained assumptions about functional form. Given this uncertainty, those unfamiliar with the literature might be surprised to learn that empirical applications typically rely on one of three functional forms—linear, semi-log, and double-log—and rarely involve rigorous specification testing. Those familiar with the literature know that

this approach is rooted in a simulation study by Cropper, Deck and McConnell (1988).[1] Their study exploits parametric assumptions about consumers' utility functions to solve for a vector of prices that clears a hypothetical housing market, using data on the structural characteristics of real homes. The simulated prices and characteristics are then used to estimate several versions of the hedonic price function, each based on a different functional form. One of their key findings is that simple functional forms (such as the linear, semi-log, and double-log) tend to convey the smallest errors in estimating MWTP for an amenity when one or more housing characteristics cannot be observed by the econometrician.

The purpose of our paper is to update and extend the CDM (1988) simulations to investigate current issues in hedonic modeling. In the 20 years that have elapsed since their study, there have been major computational advances and significant changes in the way hedonic price functions can be estimated in empirical work. This includes a variety of panel data techniques, as well as semiparametric and nonparametric methods. Furthermore, housing data in electronic formats are much more accessible and it is not uncommon for recent hedonic studies to use thousands of housing observations in an analysis. Compared to the parametric cross-sectional models considered by CDM, the newer econometric techniques present a variety of tradeoffs with respect to sample size, omitted variables, and measurement error, to name only a few issues. To date, there has been no effort to systematically evaluate the relative performance of these techniques in a controlled simulation.

In this paper the results obtained from a theoretically consistent Monte Carlo simulation that evaluates the relative performance of earlier techniques to more modern techniques are presented. Although the results presented in this paper are preliminary, three of the results not found in CDM are worth highlighting: (i) we find that adding spatial fixed effects (census tract

[1] We will refer to this paper as CDM throughout the text.

dummies) to linear models does improve their performance. This is true both when all attributes are observed, and when some attributes are unobserved, (ii) adding the spatial fixed effects to the more flexible specifications such as the quadratic and quadratic box-cox does not improve their performance when all housing attributes are observed, however, when some housing attributes are unobserved, the spatial fixed effects significantly improves the flexible specification performance as well, and (iii) increasing the sample size from CDM's 200 observations to a sample size of 2000 (which is more representative of modern applications) changes the relative performance of different specifications. We think that these results can provide additional guidance for empirical researchers. Our intention is to build off of our simulation framework in the future to extend the analysis to a panel data environment, to analyze alternative choices for utility functions and to complete nonparametric and semi-parametric specifications with larger sample sizes.

The paper proceeds as follows. In section 2 we provide a brief review of the functional form issues in hedonic property value literature. Section 3 describes our simulation framework. Section 4 highlights the results and section 5 concludes.

2. A Brief Review of Functional Form in the Hedonic Property Value Literature

In his seminal 1974 paper, Sherwin Rosen strengthened the economic foundations of the hedonic method by demonstrating that the functional relationship between the price of a differentiated product and its attributes can be interpreted as an equilibrium outcome from the interactions between all the buyers and sellers in a market. Under the assumptions of his model, regressing product prices on their attributes can reveal consumers' willingness-to-pay for a marginal change in a continuous attribute of a differentiated product (MWTP). This result has been applied to

housing markets to evaluate the welfare implications of changes in public goods and environmental amenities such as school quality (Black, 1999), air quality (Chay and Greenstone, 2005), water quality (Leggett and Bockstael, 2000), cancer risk (Davis, 2004), open space (Irwin, 2002), hazardous waste (McCluskey and Rausser, 2003), and airport noise (Pope, 2008) to name only a few. In all of these studies, estimates for welfare measures and their policy implications rely on the maintained assumption that the econometrician has correctly specified the true form of the equilibrium price function.

In Rosen's theoretical model, the form of the equilibrium price function depends on the underlying distributions of preferences and technology. Under specific parametric assumptions about these latent distributions, such as Tinbergen's (1959) linear-normal model, the equilibrium price function can take a convenient closed form. In general however, it is nonlinear without a closed-form solution. Moreover, Ekeland et al. (2004) demonstrate that nonlinearity is a generic property of the hedonic price function. This means a linear function form would be a special case in the sense that marginal perturbations to the underlying distributions of preferences and technology can produce large deviations from linearity.

While theory suggests the equilibrium price function is nonlinear, most empirical studies treat linearity as a maintained assumption. This practice is often justified by citing Cropper, Deck, and McConnell's (1988) Monte Carlo analysis of how the accuracy in predicting MWTP varies across competing functional form assumptions. The distinguishing feature of their study (henceforth CDM) is that it is theoretically consistent. They use Wheaton's (1974) linear programming algorithm to solve for an equilibrium vector of housing prices under specific assumptions about the parametric form of utility, the distribution of preferences, and the supply of housing. This allows them to compare the "true" MWTP for each housing characteristic (e.g.

bedrooms, square feet) with the econometric predictions made by each of six functional forms: *linear*, *semi-log*, *double-log*, *quadratic*, *linear Box-Cox*, and *quadratic Box-Cox*. When every housing characteristic which enters the utility function is included as an explanatory variable in the hedonic regression, the linear Box-Cox and quadratic Box-Cox produce the lowest mean percentage error in estimating MWTP. This result changes when one of the characteristics is unobserved or replaced by a proxy. In this case, the more parsimonious functional forms—linear, semi-log, double-log, and linear Box-Cox—are the ones which perform the best.

The results from CDM's "omitted variable" specification have guided the subsequent empirical literature. This is at least partly due to widespread concern about omitted variable bias in property value studies. In many recently published applications, authors' adopt a linear or a linear Box-Cox form to represent the equilibrium price function with little or no discussion of specification testing and the potential for bias.

In the 20 years that have passed since CDM's study, advances in microeconometric methods, together with the increasing availability of spatially delineated micro data, have changed the way we estimate hedonic models. Modern property value studies use econometric techniques and descriptions for the spatial landscape which differ in many ways from CDM's simulations. To document these differences, we reviewed the 110 studies published between November 1988 and April 2008 which cite CDM according to the *Social Science Citation Index* (SSCI). In addition to empirical property value studies, this set of papers includes theoretical work and applications to markets for labor, breakfast cereal, fruit, automobiles, herbicides, knitted garments, appliances, collectable coins, television, fish, forestry, and agricultural land. Narrowing the focus to residential property value studies decreased the size of our sample to 61

papers.[2] Table 1 compares the features of these studies to CDM.[3]

The influence of CDM on the choice of functional form is immediately apparent. More than three quarters of the studies in our SSCI sample (47) rely on one of the three linear functional forms: linear, semi-log, and double-log. Most of the others use a linear Box-Cox. Meanwhile, compared to CDM, the typical hedonic study uses more dummy variables, a larger sample size, a broader definition for the housing market, and explicitly controls for variation in unobserved attributes across space and time.

As data on individual housing transactions have become increasingly available, sample sizes have increased. The median number of observations in hedonic studies published during the past ten years (2,066) more than tripled from the previous ten year period (593) which was nearly triple the number of observations used by CDM (200). As sample sizes have grown, so have the geographic and temporal boundaries used to define a housing market. CDM used data on homes sold in Baltimore City and Baltimore County in 1977-78. In comparison, 32 of the 61 papers in our sample use data from multiple cites or counties, and 34 use sales data over more than two years. Gayer et al. (2000) provide a representative example. They use data on approximately 17 thousand homes sold in the greater Grand Rapids, MI area over a six year period.

Over the past 20 years, the literature has also evolved to address omitted variables directly. More than half the studies in the SSCI sample (35) use dummies to absorb the effect of unobserved amenities that vary between cities or between "neighborhoods" within a city (e.g. census tracts, school districts). A smaller set of papers (7) use spatial econometrics to impose

[2] A complete list of these papers is provided in a supplemental appendix available from the authors upon request.
[3] Many of these studies report the results from multiple econometric models. We focus on the model which the authors identify as their main specification. If the authors do not identify a main specification, we focus on the model which produces the results which enter their discussion of policy implications and/or conclusions.

more structure on the spatial relationship between unobserved variables and housing prices. Perhaps most importantly, researchers are increasingly using fixed effects, first difference, and difference-in-difference estimators to exploit changes in amenities over time as an identification strategy. Of the 15 studies which exploit the panel structure of their data for identification, 11 were published since 2000. These studies are often able to make a convincing argument that changes in housing prices are *caused* by changes in the amenity of interest. Moreover, the availability of data on repeated sales of individual homes provides a way to fully purge time-constant omitted variables (e.g. Davis, 2004). None of these new strategies for addressing omitted variables were considered by CDM. The bottom line is that the empirical hedonic literature which routinely invokes the results from CDM has evolved to the point where it bears little resemblance to their original study.

Three features of the hedonic literature suggest to us that it is time to revisit the issue of functional form. First is the emergence of new techniques for addressing omitted variables. To the best of our knowledge, there is no existing evidence on the relative performance of the different techniques within a theoretically consistent simulation framework. We hypothesize that it may be possible to increase the accuracy of estimates for MWTP by extending nonlinear estimators to include spatial dummy variables, spatial error corrections, and panel data.

Second, the increase in sample size documented in table 1 highlights the relatively small number of homes (200) used by CDM in their Monte Carlo simulations. Larger sample sizes are needed to implement nonparametric estimators as well as some of the techniques for addressing omitted variables. Furthermore, in a simulated differentiated product market, the equilibrium difference between true MWTP and true marginal implicit prices will depend on the extent of

discreteness in the choice set.[4] Banzhaf (2003) provides preliminary evidence that this difference has the potential to be economically important. Holding the number of covariates constant, increasing the sample size will tend to provide a more continuous choice set and decrease the divergence between MWTP and marginal implicit prices. It is difficult to anticipate whether this will alter conclusions about the relative performance of different functional forms.

Finally, the econometric literature on nonlinear estimation has moved beyond the Box-Cox specifications considered by CDM.[5] Empirical evidence suggests that newer semiparametric and nonparametric methods dominate the linear Box-Cox in terms of in-sample and out-of-sample ability to predict prices (e.g. Gençay and Yang, 1996; Bin, 2004). Will these methods also outperform the linear Box-Cox in estimating MWTP? Or will the additional nonlinearity lead to abysmal performance in the presence of omitted variables, as CDM observed for the quadratic Box-Cox? It is important to answer these questions as semi and nonparametric methods are beginning to be used to address important policy issues, such as racial segregation (Bajari and Kahn, 2005).

3. Simulation Framework

In order to investigate how our ability to accurately estimate MWTP depends on the size of the simulation, methods for mitigating omitted variable bias, and the use of semi and nonparametric estimators, we follow CDM in developing a theoretically consistent Monte Carlo simulation. After briefly reviewing the equilibrium concept, we describe the numerical algorithm we use to solve for equilibria and then summarize the features of the data we use to simulate the housing

[4] The first order conditions of Rosen's hedonic model require that consumers are able to set their marginal rates of substitution equal to the ratio of marginal implicit prices. If consumers are not free to choose from a continuum of choices, this condition cannot be met, creating a divergence between MWTP and marginal implicit prices. For a formal discussion see Bajari and Benkard (2005), Bayer et al. (2007), or Kuminoff (2008).
[5] See Parmeter (2006) for a review of hedonic applications of nonlinear estimators.

market in Wake County, North Carolina.

3.1. Characterizing a Locational Equilibrium

Suppose the availability of housing and amenities varies across an urban landscape and that each household chooses the particular home which provides its preferred bundle of goods, given its preferences, income, and relative prices. This problem can be formalized using the characteristics approach to consumer theory (Lancaster, 1966). Let $j = 1,....,J$ homes be defined over a vector of characteristics, x_j. This includes structural characteristics of the home, such as the number of bedrooms, the number of bathrooms, square feet, and lot size, as well as amenities, such as crime, school quality, air quality and proximity to open space. A household's utility depends on the characteristics of housing and amenities at its location and on its consumption of a composite numeraire, *c*. Households are heterogeneous. They differ in their income, y, and in their preferences, α. Let the population of households be indexed from $i = 1,....,N$. Each household is assumed to choose a specific house and a quantity of *c* that maximize its utility subject to a budget constraint:

$$\max_{j,c} \; U\left(x_j, c; \alpha\right) \;\; subject \;\; to \;\; y = \; c + p_j . \tag{1}$$

In the budget constraint, the price of the numeraire is normalized to one, and p_j represents annualized expenditures on house *j*.

A locational equilibrium is achieved when every household occupies its utility-maximizing location and nobody wants to move, given housing prices and their exogenously

determined characteristics.[6] In order to define this concept more formally, let b_{ij} denote household *i*'s bid for the jth home, and let A_{ij} be an assignment indicator where $A_{ij} = 1$ if and only if household *i* occupies that home. Then a locational equilibrium can be defined as follows:

$$b_{ij} = \max_i \left\{ b_{ij} \right\} \quad iff \quad A_{ij} = 1, \tag{2}$$

$$\sum_i A_{ij} = \sum_j A_{ij} = 1. \tag{3}$$

In words, each household occupies exactly one home, for which it has the maximum bid.[7]

In the context of Rosen's (1974) hedonic model, bids can be expressed as a function of housing characteristics and preferences. To see this, let $\tilde{u}$ be some reference level of utility, and consider an indifference surface over which x and *c* vary, while $\tilde{u}$ stays the same: $\tilde{u} = U(x, c; \alpha)$. Assuming utility is monotonically increasing in *c*, the function can be inverted to solve for *c*.

$$c = U^{-1}(x, \tilde{u}; \alpha). \tag{4}$$

Inserting (4) into the budget constraint and rearranging terms allows a household's maximum willingness-to-pay for a home to be expressed as a function of its characteristics and the household's income, preferences, and utility.

$$b = y - U^{-1}(\tilde{u}, x; \alpha). \tag{5}$$

[6] See Bayer and Timmins (2007) for a discussion of equilibria and estimation in location choice model with endogenously determined public goods.
[7] Equations (2)-(3) are equivalent to the equilibrium concept defined in equations (2)-(4) of CDM.

This is Rosen's (1974) bid function. It can be used to solve for a locational equilibrium, given a parametric specification for the utility function, information on preferences and income, and data on housing characteristics.

3.2. Numerical Approach to Solving for a Locational Equilibrium

Kuminoff and Jarrah (2008) develop an Iterative Bidding Algorithm (IBA) which uses Rosen's bid function to solve for a locational equilibrium. The IBA iterates over a series of hypothetical second-price auctions for individual homes until subsequent bidding has no effect on prices or the assignment of people to homes; i.e. until equations (2)-(3) are simultaneously satisfied.

The algorithm begins by assigning each household a reference level of utility. This can be used together with data on x and the joint distribution $f(\alpha, y)$ to solve for each household's bid for each home. Consider the first home. The IBA uses (5) to solve for the distribution of bids and assigns the household with the maximum bid to live there. However, the household does not pay its full bid. It pays the second highest bid plus a marginal increment, $\varepsilon > 0$. The resulting price, p_1, is then used to update the household's reference utility. Since it pays less than its maximum bid, utility must increase. The same process is used to update the prices of homes 2 through *J*, and the reference utility of the households who are assigned to live there.

To provide a formal description of the algorithm, we use superscripts to rank households according to their bids. For example, $b_j^1 \geq b_j^2 \geq ... \geq b_j^{J-1} \geq b_j^J$ depicts a ranking of households according to their bids for the jth home. Household-specific subscripts are suppressed to preserve generality. Using this notation, the first iteration of the IBA is shown as (6).

$$
\begin{array}{lll}
1. & p_1 = \min\{b_1^1, b_1^2 + \varepsilon\}, & \tilde{u}^1 = U(x_1, y^1 - p_1; \alpha^1). \\
2. & p_2 = \min\{b_2^1, b_2^2 + \varepsilon\}, & \tilde{u}^1 = U(x_2, y^1 - p_2; \alpha^1). \\
 & \quad . & \quad . \\
 & \quad . & \quad . \\
 & \quad . & \quad . \\
J. & p_J = \min\{b_J^1, b_J^2 + \varepsilon\}, & \tilde{u}^1 = U(x_J, y^1 - p_J; \alpha^1).
\end{array}
\tag{6}
$$

Two features of (6) are notable. First, the minimum operator is used to ensure that households do not pay more than their bid in the special case where the top two bids are identical. Second, the 1 superscript on $\tilde{u}^1$, y^1, and α^1 may denote a different household on each of the *J* steps of (6) because the bid ranking will vary across homes due to heterogeneity in reference utility and preferences for housing characteristics.

The price vector that results from the first iteration of (6) need not support a locational equilibrium. For example, the household who was assigned to home 1 may have been subsequently assigned to another home. The corresponding increase in utility will decrease its bid for the first home so that p_1 must decrease in order to clear the market. This decrease in p_1 may necessitate decreases in the prices of other homes. Therefore, the IBA continues to iterate over (6) until the price vector converges, signaling the market has cleared.

Unlike the linear programming algorithm used in the simulations reported by CDM and Banzhaf (2003), the IBA does not constrain households to be uniquely assigned to homes on any given iteration. Nevertheless, if the algorithm converges it must converge to a locational equilibrium. Moreover, if a unique equilibrium exists, the algorithm will converge to it. Kuminoff and Jarrah (2008) provide proofs, examples, and a discussion of computational

issues.[8] We use their algorithm here to simulate hedonic equilibria in Wake County, North Carolina.

3.3. Simulating Hedonic Equilibria in Wake County's Market for Housing

Wake County, North Carolina is the geographic setting for the analysis in this study. According to the 2000 census there were approximately 628 thousand people living in the county in 1999. About 72 percent of the population is white, 20 percent black, and 6 percent Hispanic/Latino. The median household income in 1999 was approximately $55,000. The largest city in the county is Raleigh with a reported population of 276 thousand as of 1999. Most of the remaining population lives in 12 satellite municipalities in the county, with the biggest of these being the town of Cary. The census name for the metropolitan area is the Raleigh-Cary NC metropolitan area.

Wake County provides an ideal setting for a simulation exercise aimed at understanding some of the empirical concerns of the hedonic literature. First, because of the population and number houses in the county, there are thousands of housing transactions that occur in this housing market every year. Therefore, housing data from this area will likely accommodate the needs of our simulation exercise to use a sample size that is an order of magnitude larger than that used in CDM. The area is also well suited for our analysis because of differences in amenities across municipalities within the county. This naturally gives rise to a need for implementing some of the modern methods of controlling for omitted variable bias that are used in some of the simulations.

[8] While the IBA is based on the same equilibrium concept as the algorithm used by CDM, it has two desirable properties which set it apart and make it more appropriate for our study. First, it can be demonstrated that the IBA will necessarily converge to an equilibrium, if one exists. Second, the IBA avoids the need to store large assignment matrices in the computer's memory, enabling us to increase the size of our simulated market to the point where it is reasonable to include spatial fixed effects and to use nonparametric methods.

Solving for a hedonic equilibrium using the Iterative Bidding Algorithm described earlier, requires defining the stock of housing and the joint distribution of income and preferences. The stock of housing is defined using actual housing data originally obtained from the Wake County Revenue Department. The dataset has been used in a variety of contexts including; Fulcher (2002), Pope (2008), Phaneuf et al. (forthcoming) and Pope (forthcoming). The data spans the years 1992 to 2000 and contains approximately 104 thousand observations of houses that transacted over this time frame. This dataset is much more complete than most datasets used in typical hedonic analyses because of detailed information about the square feet of various components of the house (i.e. garages, decks, basements and attics). However, to keep the simulation exercise as realistic as possible, we limit the variables to those found in typical hedonic analyses. Furthermore, although we have information on the square feet of garages and the total number of fireplaces, we convert these two variables to dummy variables that indicate whether or not a home has a garage or a fireplace. This is the most common way in the literature for information on these two housing characteristics to enter into a hedonic regression.

Table 2 provides summary statistics of the housing prices (our dependent variable) and 11 housing characteristics (our independent variables) used in our analysis.[9] In our simlulations, each of these 11 variables enter into utility. Note that this is approximately the same number of characteristics used in CDM.[10] The average house in the dataset sells for approximately 201 thousand dollars, has 2.5 baths, is on .5 acres, does not have a garage, has a fireplace, has 1900

[9] We converted housing prices to rents for our simulations using the formula from Poterba (1992). Poterba's formula is: $R = [(1-\tau)(i-\tau_p) + r + m + \delta - \pi]P$. Where for Wake County, τ is the owner's marginal tax rate and is equal to 15% according to Walsh (2007), τ_p is the property tax rate and is 0.95% according to Wake county, i is the interest rate and averages 7.76% over the 1992-2000 time period according to information reported by Freddie Mac, r is the risk premium set to 4% according to Poterba (1992), m is maintenance set to 2% according to Poterba (1992), δ is depreciation set to 2% according to Poterba (1992) and π is land appreciation rate set to 3.19% using the average of the BLS Housing Price Index over the 1992 to 2000 time frame.

[10] In the simulations we use 11 characteristics for the scenarios where all housing attributes are observed whereas CDM used 12 characteristics.

square feet of heated living space, is about 10 years old, is located in a census tract where median household income is 68 thousand dollars, commute time to work is on average 23 minutes, 27 percent of people in the census tract are under 18, is 4 miles from the nearest park larger than 70 acres and is 8 miles from one of 4 major shopping areas in the county. Table 3 presents the correlation coefficients for the independent variables. The highest correlations occur between the "nearest shopping center" variable and the "median time to work" (0.77) and "nearest park" (0.73) variables. "Main heated living area" is also highly correlated with "garage" (0.67) and "bathrooms" (0.65).

The data used in our simulations also included geographic information for each home. Variables that related each home to its corresponding census tract and block group were included along with the latitude and longitude of each home. These variables do not enter utility directly, but are used to control for spatially delineated omitted variables in some of our simulation scenarios. Figure 1 shows census tracts in the county in relation to the latitude and longitude points of each home in our dataset. Notice how homes are concentrated in the center of the county where the city of Raleigh is located.

We represent each household's utility using the Cobb-Douglas specification in (7), where X_j is a vector of continuous housing characteristics (e.g. # bathrooms, age) and D_j is a vector containing dummy variables indicating whether the home has a fireplace and a garage.

$$U_{ij} = \ln(c) + \alpha_i \ln X_j + \beta_i D_j \,. \tag{7}$$

Preferences for housing characteristics are assumed to be independent of income and gamma distributed. Selecting a gamma distribution recognizes that the distribution of preferences may not be symmetric about the mean. This makes it easier to calibrate the simulation to

approximately reproduce the actual distribution of housing prices in Wake County. The distribution of household income was defined using data from the 2000 *Census of Population and Housing*, which reports the number of households with income in each of 16 bins.

The price data for our Monte Carlo evaluation of functional form are generated by using the Iterative Bidding Algorithm to solve for 100 hedonic equilibria, using two different sample sizes, N=200 and N=2000. On each Monte Carlo replication, households are randomly drawn from the nonparametric Census income distribution under the assumption that people are uniformly distributed within each bin.[11] Then, given a random sample of homes and a random sample of income, a quasi-Newton algorithm is used to solve for values of the gamma shape and scale parameters which minimize the distance between predicted and observed equilibrium housing prices.

Figure 1 contrasts the difference between the predicted and observed distributions of prices on a representative Monte Carlo replication. For example, the solid line in panel A represents the empirical cumulative distribution function of actual prices for 200 homes in Wake County.[12] The dashed line represents the equilibrium prices assigned to those homes in our simulation. While the predicted prices for some individual homes differ considerably from their actual values, the simulation clearly reflects the general price trend in our data. This is reinforced by the close match between the corresponding simulated and empirical probability density functions in panel B. Panels C and D illustrate that these results do not change when we increase the sample size to 2000. Overall, our simulated equilibria appear to provide a reasonable approximation to the observable features of the housing market in Wake County.

[11] The lowest income bin (y<$10,000) was dropped under the assumption that households in this category are retired or purchasing housing out of savings. The top income bin (y>$200,000) was truncated at $300,000 for the purposes of the simulation.

[12] Recall that these are annualized housing prices. Converting them back to actual housing prices would require multiplying by 1/.1222.

4. Results

The data on housing characteristics are combined with the simulated equilibrium prices generated on each of our 200 Monte Carlo replications to estimate 20 specifications for the hedonic price function. We begin by considering the six functional forms from CDM's original study: *linear*, *semi-log*, *double-log*, *linear Box-Cox*, *quadratic*, and *quadratic Box-Cox*. The first four have dominated the empirical hedonic literature for the past two decades (table 1). Our next twelve specifications simply add dummy variables for Census tracts, and then Census blocks, to the six functional forms from CDM. This allows us to evaluate how the increasingly common practice of adding spatial fixed effects to the hedonic price function influences performance in the presence of omitted variables. Next, to compare spatial fixed effects with the parametric literature on spatial econometrics, we estimate a spatial error model based on a contiguity matrix for homes within 0.13 degrees of one another. Kim et al. (2003) provides a nice discussion of this type of spatial regression model in a hedonic application.

Our final econometric specification is a fully nonparametric model which nests all of the parametric specifications from CDM and many popular semi-parametric specifications (Li and Racine [2007]). Nonparametric methods are robust to model misspecification, making them well suited to hedonic estimation. Recent advances in these techniques have increased their appeal in applied settings. Specifically, Li and Racine (2007) document several ways in which the curse of dimensionality can be reduced by accounting for variables which are inherently discrete and by removing irrelevant variables from the outset. These features allow standard nonparametric estimation at smaller sample sizes than commonly advocated.

Equation (8) defines the model, where x_i represent the continuous covariates and z_i represents discrete covariates.

$$p_i = m(z_i, x_i) + \varepsilon_i \,, \tag{8}$$

Four of our covariates are discrete: *number of bathrooms*, *garage*, *fireplace*, and *age*. We add the unique identification number of each census tract as a fifth discrete covariate. This serves to control for unobserved characteristics which are correlated with tract id number.

We estimate the model in (8) using the Li-Racine generalized kernel local linear approach with bandwidths selected via least squares cross validation (Li and Racine, 2007). This method has three features that make it well suited to hedonic estimation. First, a generalized kernel estimator is capable of distinguishing between discrete and continuous covariates. This distinction is important because, as Racine and Li (2004) observe, the convergence rates on the bandwidths depend only on the number of *continuous* covariates. Second, least squares cross validation has been shown to automatically remove irrelevant variables (Hall et. al. 2007). This occurs as the bandwidths of the irrelevant variables are set to their theoretical upper bounds.[13] While all of the variables in our model are relevant in the sense that they enter the utility function, their relevance in explaining equilibrium prices is unknown a priori as is often the case for a subset of the covariates in empirical hedonic studies. Finally, the local linear approach estimates the unknown function and its derivatives simultaneously.

An important caveat to our nonparametric approach is that large sample sizes are needed to obtain desirable statistical properties for the gradient of $m(\cdot)$. The traditional convergence rate

[13] In finite samples this amounts to selecting a bandwidth that is larger than two or three standard deviations of the continuous variables.

of the nonparametric estimator of the unknown function $m(\cdot)$ is $O_p\left(\eta_1 + \sqrt{\eta_2}\right)$, where $\eta_1 = \sum_{s=1}^{q} h_s^2$, $\eta_2 = \left(nh_1 \cdots h_q\right)^{-1}$, and h_s is the bandwidth associated with the sth continuous covariate. In comparison, the convergence rate for the sth derivative is $O_p\left(\eta_1 + \sqrt{\eta_2} h_s^{-1}\right)$. The presence of h_s^{-1} in the convergence rate slows the asymptotic speed with which estimates for the derivatives of $m(\cdot)$ converge to their true values.[14]

To evaluate the relative performance of the 20 different functional forms, we first calculate the difference between every household's MWTP for each housing characteristic and the corresponding partial derivative of the hedonic price function, $P(x)$. Equation (9) defines this difference for household *i*'s valuation of characteristic *k* on Monte Carlo replication *r*.

$$e_{ikr} = \partial P_r\left(x_i\right) / \partial x_k - MWTP_{jkr} . \tag{9}$$

We follow CDM by using (9) to construct summary statistics for the distribution of errors in estimating MWTP for the population of households. Equation (10) defines the normalized mean (β_{kr}) and standard deviation (S_{kr}) of the errors for each attribute on a given replication.

$$\beta_{kr} = \frac{\bar{e}_{kr}}{N^{-1}\sum_i MWTP_{ikr}}, \qquad S_{kr} = \frac{s_{kr}}{N^{-1}\sum_i MWTP_{ikr}}, \qquad k = 1,\ldots,K \tag{10}$$

The normalized mean and standard deviation are simply the mean ($\bar{e}_{kr}$) and standard deviation (s_{kr}) of the error from equation (9), divided by the average MWTP for characteristic *k*.

[14] An alternative approach would be to use a local cubic estimator, which would improve the convergence of the derivative estimator at the expense of requiring more smoothness of the hedonic function.

Like CDM, our simulation is designed to evaluate the potential for omitted variables to contaminate econometric estimates for MWTP. Yet we focus on a different class of omitted variables. CDM omit two structural characteristics—lot size and the number of rooms. In the twenty years since their study, data on structural characteristics have become readily available. Detailed information on the characteristics of each home (including lot size and the number of rooms) are virtually always included in the "assessor" property value databases which are now used in most hedonic studies.[15] Nevertheless, concern about omitted variable bias has intensified. Since most studies seek to estimate the MWTP for spatially delineated amenities (e.g. air quality, flood risk, airport noise, proximity to registered sex offenders) concern about omitted variable bias has shifted to unobserved features of neighborhoods. For example, suppose we seek to measure the willingness-to-pay for a marginal increase in the distance of a home from a landfill. If homeowners care about crime rates, and landfills tend to be located in high-crime areas, failing to include crime rates in the price function may artificially inflate estimates for MWTP.[16]

In our omitted variable scenario, the econometrician observes only one of the spatially delineated attributes in table 1, *nearest park*. That is, *distance to the nearest shopping center* is omitted along with the three Census block variables (*median household income*, *% under 18*, and *time-to-work*). Without any form of correction, omitting these four variables should artificially inflate estimates for the MWTP for distance to the nearest park. Table 3 illustrates that, all else constant, moving to a home located further from parkland generally means moving to a lower-

[15] County assessors are often required to keep detailed records of the structural characteristics and transaction price of every home sold in the county for tax purposes. This public information is collected by several commercial vendors, including *Dataquick* and *TransAmerica Intellitech*, who package it in electronic databases for sale to researchers and marketing firms.

[16] Spatial dummy variables are often included in the price function with the intention of "absorbing" the price effects of unobserved neighborhood characteristics.

income community, experiencing a longer commute, and increasing the distance to shopping centers, all of which decrease utility.

4.1. Comparison of Basic Results to CDM

Table 4 summarizes our basic results for the first six functional forms, providing a quick comparison to CDM. The summary measures $|\beta_k|$, S_k, and β_k are calculated over all 100 Monte Carlo replications and over the seven housing characteristics which enter every econometric specification (*bathrooms*, *acreage*, *garage*, *fireplace*, *heated area*, *age*, and *park*). For example, when all 11 characteristics are observed and a semi-log model is used in the simulation with 200 homes, estimates of the MWTP for individual characteristics differ from the true MWTP in absolute terms by 54% on average, and the maximum difference for any characteristic is 93%. Moving from left to right in the table, the econometric flexibility increases from a simple linear specification to a quadratic model of Box-Cox transformed variables. As in CDM, when every attribute which influences households' location choices is used in the estimation, the quadratic Box-Cox model outperforms all other functional forms.

Increasing the size of the simulated market from 200 homes to 2000 homes has two important consequences. First, it changes the relative performance of some of the specifications. The double-log specification has the fourth largest maximum $|\beta_k|$ when N=200 and the second largest when N=2000. Similarly, the quadratic model has by far the largest S_k when N=200, and an intermediate value when N=2000. The second important consequence of increasing the market size is that it decreases the magnitude of the capitalization bias reflected in the average values for β_k. That is, when households with heterogeneous preferences face a discrete set of

choices they may be unable to sort themselves according to every characteristic, as Banzhaf (2003) observed. As a result, the willingness-to-pay for some housing attributes may be less than fully capitalized into equilibrium prices.[17] Indeed, the fourth row of the table illustrates that the MWTP for housing characteristics is systematically underestimated. Notice that the magnitude of this effect decreases when the market size is increased to 2000 in simulation #2. There are two reasons. Increasing *N* tends to "fill in" the space of housing characteristics, which increases the opportunity for households to sort themselves across the urban landscape according to their preferences. Likewise, increasing *N* tends to "fill in" preference space, which increases the competition between households, requiring their equilibrium bids to more fully reflect their maximum willingness-to-pay.

When four variables are omitted in simulation #3, the average normalized error and the maximum error increase in every econometric specification except the double-log. As in CDM, the quadratic Box-Cox appears to be affected the most. While the direction of this effect is consistent with CDM, the magnitude is noticeably smaller. This difference underscores the importance of repeating our simulations for different data sets and alternative specifications for the parametric form of the utility function—two of the next steps in this research.

To compare the performance of the six specifications from CDM with advances in nonparametric methods and new strategies for addressing omitted variables, we report the full set of results from all 20 econometric specifications in tables 5 through 8.

4.2. New Strategies for Dealing with Omitted Variable Bias: Cross-Section

Columns (7) through (13) in Table 5 present the specifications that utilize two new strategies for dealing with omitted variable bias for the case where all housing attributes are observed and

[17] See Starrett (1981) for a general discussion.

there are 200 observations. Columns (7) through (12) include 70-75 census tract dummies in the six base specifications and column (13) runs the semi-log functional form with a spatial error correction. Focusing on average $|\beta_k|$, it can be seen that the linear, semi-log, double-log and the linear box-cox models all make modest improvements whereas the quadratic and box-cox quadratic models perform more poorly than without census tract controls. For example, the semi-log model goes from 0.56 to 0.52. The spatial error model performs about the same as its semi-log counterpart in column (2).

Now focusing on the corresponding columns in Table 6, it can be seen that the improvements in average $|\beta_k|$ is even more substantial for the linear, semi-log, double-log and the linear box-cox models when some housing attributes are unobserved and there are 200 observations. For example, the linear model improves from 0.67 to 0.56 when census tract controls are added to the specification. Furthermore, with unobserved attributes, the census tract controls improve the performance of the more flexible quadratic and quadratic box-cox specifications. The quadratic goes from 0.51 to 0.43 and the quadratic box-cox goes from 0.38 to 0.30. When all attributes were observed in Table 5, the quadratic model performed better than all but the quadratic box-cox model. However in Table 6, much like what was reported by CDM, when some housing attributes are unobserved the quadratic performs worse than two of its linear counterparts. However, with the inclusion of census tract dummies the quadratic model closes the performance gap with the two linear models actually barely outperforming one of them. The inclusion of the census tract dummies into the quadratic box-cox model dramatically improves its performance such that it appears to be almost as accurate as when all housing attributes are observed.

Tables 7 and 8 correspond with Tables 5 and 6, but now the number of observations has

been increased to 2000. Many of the conclusions we drew from tables 5 and 6 on the value of adding the spatial dummies holds when the sample size increases as well. One important difference is that the relative performance of some of the specifications has changed. For example, the double-log model performs more poorly relative to some of its counterparts when the sample size is increased to 2000. In Table 8 we also tried the same set of 6 base specifications with even more disaggregated spatial controls. We included census bock group dummies and found a very modest improvement beyond previous specifications.

4.3. Nonparametric Estimator

Of the 14 specifications summarized in tables 5 and 6, the nonparametric estimator has the largest average error (measured in absolute terms) and is also the most sensitive to the presence of omitted variables. It seems likely that the model's relatively poor performance simply reflects the small sample size. We are in the process of repeating the nonparametric estimation for the N=2000 scenario, where we would expect the slow convergence speed for estimates of the gradient of the hedonic price function to pose less of a problem. Finally, note that despite the small sample size in tables 5 and 6, the nonparametric estimator outperforms the linear and semi-log models in its ability to recover the MWTP for distance to the nearest park.

5. **Conclusions**

The hedonic pricing model is widely accepted as a method for estimating the marginal willingness to pay for spatially delineated amenities. Empirical applications typically rely on one of three functional forms—linear, semi-log, and double-log—and rarely involve rigorous

specification testing. This phenomenon is largely due to an influential simulation study by Cropper, Deck and McConnell (1988) that found among other things that simpler linear specifications outperformed more flexible functional forms in the face of omitted variables. In the 20 years that have elapsed since their study, there have been major computational advances and significant changes in the way hedonic price functions can be estimated. The purpose of our paper is to update and extend the CDM (1988) simulations to investigate current issues in hedonic modeling.

Three preliminary results obtained from our theoretically consistent Monte Carlo simulation have been highlighted in this paper: (i) we find that adding spatial fixed effects (census tract dummies) to linear models does improve their performance. This is true both when all attributes are observed, and when some attributes are unobserved, (ii) adding the spatial fixed effects to the more flexible specifications such as the quadratic and quadratic box-cox does not improve their performance when all housing attributes are observed, however, when some housing attributes are unobserved, the spatial fixed effects significantly improves their performance as well, and (iii) increasing the sample size from CDM's 200 observations to a sample size of 2000 (which is more representative of modern applications) changes the relative performance of different specifications.

Our intention for future research is to extend our analysis in three directions. First we will study the effect of alternative choices of utility functions on the results of our simulations. Second we will simulate a panel of housing data to explore issues of hedonic functional form in quasi-experimental analyses that identify an effect over time. Finally we will run additional spatial regression, parametric and nonparametric specifications for larger sample sizes. We think that the results presented in this paper along with results derived from the proposed extensions to

this work can provide additional guidance for empirical specification of hedonic price functions and will fill an important gap in the literature.

Table 1: Features of Empirical Hedonic Applications: 1998-2008

Total # studies		CDM	SSCI (61)
Functional form	# using lin-lin, log-lin, log-log		47
	# using Box-Cox		12
Dummy Variables	Mean share of covariates which are 0/1	17%	36%
Sample Size	Median # observations	200	1,679
	published in 1989-1998		593
	published in 1999-2008		2,066
	Distribution of studies by # observations		
	0 to 200	1	5
	201 to 500		6
	501 to 1,000		16
	1,001 to 10,000		25
	more than 10,000		9
Housing Market	Geography (#)		
	smaller than a city		5
	city or county	1	24
	multiple cities or counties		28
	nation		4
	Time Period (#)		
	0 to 1 year		16
	1 to 2 years	1	11
	2 to 5 years		14
	5 to 10 years		11
	more than 10 years		9
Space and Time	# with spatial error or spatial lag structure		7
	# exploiting panel structure of data		15
	# with time dummies or time trend		23
	day		3
	month		3
	quarter		2
	year		15
	# with spatial dummies		35
	neighborhood		13
	city or county		20
	region		2

Table 2: Summary Statistics for Wake County, North Carolina

type	Variable	Units	Mean	Std.	Min	Max
price	price	$1,000	201	105	16	2976
structural	bathrooms	#	2.50	0.76	1.00	10.50
structural	acreage	#	0.50	0.93	0.01	97.52
structural	garage	dummy	0.29	0.26	0.00	1.00
structural	fireplace	dummy	0.91	0.36	0.00	1.00
structural	main heated living area	sqft (1000)	1.93	0.73	0.40	9.08
structural	age	years	10.38	15.05	1.00	99.00
block	median household income	$1,000	67.87	21.30	8.32	146.76
block	median time to work	minutes	22.71	4.49	7.00	37.00
block	% under 18	%	26.77	5.18	2.15	49.84
amenity	nearest park	miles	4.34	2.84	0.41	18.59
amenity	nearest shopping center	miles	7.86	4.76	0.39	26.07

Table 3: Correlation Coefficients for Covariates

type	Variable	bath-rooms	acreage	garage	fireplace	main heated living area	age	median household income	median time to work	% under 18	nearest park	nearest shopping center
structural	bathrooms	1.00										
structural	acreage	0.07	1.00									
structural	garage	0.53	0.07	1.00								
structural	fireplace	0.34	0.04	0.23	1.00							
structural	main heated living area	0.65	0.14	0.67	0.32	1.00						
structural	age	-0.39	0.04	-0.43	-0.18	-0.28	1.00					
block	median household income	0.50	0.06	0.50	0.27	0.57	-0.29	1.00				
block	median time to work	-0.04	0.14	0.08	-0.01	-0.07	-0.37	-0.08	1.00			
block	% under 18	0.22	0.07	0.32	0.07	0.23	-0.46	0.50	0.47	1.00		
amenity	nearest park	-0.11	0.14	-0.07	-0.08	-0.14	-0.13	-0.18	0.54	0.20	1.00	
amenity	nearest shopping center	-0.10	0.18	-0.02	-0.06	-0.11	-0.28	-0.18	0.77	0.37	0.73	1.00

Table 4: Preliminary Comparison to CDM

Simulation	Criterion	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
				No Omitted Variables, N=200			
#1	Maximum $\lvert \beta_k \rvert$	0.97	0.97	0.93	1.08	0.88	0.62
	Average $\lvert \beta_k \rvert$	0.62	0.54	0.54	0.50	0.47	0.30
	Average S_k	1.44	1.44	1.46	1.48	1.94	1.40
	Average β_k	-0.46	-0.45	-0.28	-0.19	-0.43	-0.30
				No Omitted Variables, N=2000			
#2	Maximum $\lvert \beta_k \rvert$	0.98	0.98	1.38	1.56	0.90	0.76
	Average $\lvert \beta_k \rvert$	0.69	0.60	0.59	0.55	0.49	0.27
	Average S_k	1.39	1.39	1.47	1.50	1.40	1.24
	Average β_k	-0.41	-0.41	-0.19	-0.10	-0.37	-0.27
				Four Omitted Variables, N=2000			
#3	Maximum $\lvert \beta_k \rvert$	1.39	1.14	0.92	1.01	0.96	0.81
	Average $\lvert \beta_k \rvert$	0.78	0.66	0.58	0.59	0.55	0.41
	Average S_k	1.39	1.41	1.44	1.55	1.41	1.27
	Average β_k	-0.26	-0.25	-0.06	-0.02	-0.23	-0.07

Table 5: Errors in Measuring MWTP When All Housing Attributes are Observed, *N=200*

Mean Error / Mean True Price
(Standard Deviation of Error / Mean True Price)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Spatial Error Model	Non-parametric
bathrooms	-0.106	-0.223	-0.313	-0.206	-0.294	-0.117	-0.104	-0.252	-0.149	-0.140	-0.129	-0.254	-0.217	-0.974
	(1.146)	(1.144)	(1.188)	(1.224)	(2.025)	(1.374)	(1.146)	(1.141)	(1.206)	(1.213)	(3.006)	(2.161)	(1.144)	(1.176)
acreage	-0.421	-0.477	0.932	1.079	0.138	-0.191	-0.403	-0.467	0.854	0.829	0.092	-0.125	-0.481	-0.322
	(0.958)	(0.947)	(1.524)	(1.797)	(1.109)	(1.193)	(0.958)	(0.948)	(1.486)	(1.485)	(1.239)	(1.580)	(0.947)	(0.966)
garage	-0.485	-0.286	-0.479	-0.418	-0.432	-0.437	-0.445	-0.265	-0.448	-0.431	-0.299	-0.444	-0.289	-0.932
	(1.463)	(1.435)	(1.437)	(1.437)	(1.974)	(1.517)	(1.463)	(1.435)	(1.436)	(1.436)	(3.186)	(1.590)	(1.435)	(1.451)
fireplace	-0.925	-0.663	-0.774	-0.718	-0.842	-0.625	-0.897	-0.697	-0.725	-0.706	-2.549	-0.695	-0.649	-1.000
	(1.494)	(1.480)	(1.484)	(1.482)	(2.805)	(1.610)	(1.494)	(1.483)	(1.483)	(1.482)	(5.798)	(2.808)	(1.480)	(1.494)
main heated living area	0.572	0.323	-0.133	-0.163	-0.134	-0.117	0.381	0.178	-0.238	-0.224	-0.389	-0.252	0.332	1.225
	(1.090)	(1.131)	(1.157)	(1.179)	(1.866)	(1.314)	(1.090)	(1.117)	(1.148)	(1.152)	(2.914)	(2.352)	(1.133)	(1.237)
age	-0.974	-0.969	-0.608	-0.460	-0.883	-0.392	-0.949	-0.944	-0.599	-0.583	-0.862	-0.422	-0.969	-0.836
	(1.890)	(1.887)	(1.621)	(1.522)	(1.867)	(1.447)	(1.890)	(1.884)	(1.616)	(1.606)	(1.854)	(1.484)	(1.887)	(1.819)
median household income	0.039	0.103	-0.199	-0.133	-0.211	-0.158	0.001	0.113	-0.281	-0.251	-0.347	0.004	0.110	0.244
	(1.165)	(1.175)	(1.228)	(1.271)	(2.165)	(1.427)	(1.165)	(1.198)	(1.230)	(1.242)	(4.501)	(2.684)	(1.176)	(1.308)
median time to work	-0.185	-0.147	-0.407	-0.429	-0.139	-0.048	-0.204	-0.055	-0.432	-0.417	-0.274	-0.078	-0.079	-0.660
	(1.714)	(1.705)	(1.646)	(1.641)	(2.654)	(1.772)	(1.714)	(1.718)	(1.661)	(1.664)	(4.346)	(2.800)	(1.706)	(1.843)
% under 18	-1.176	-1.259	-0.681	-0.596	-0.018	-0.074	-1.032	-1.046	-0.429	-0.423	-0.568	-0.200	-1.287	-0.733
	(1.323)	(1.367)	(1.342)	(1.357)	(3.347)	(1.699)	(1.323)	(1.359)	(1.367)	(1.378)	(5.943)	(3.253)	(1.369)	(1.559)
nearest park	-0.873	-0.850	-0.558	-0.456	-0.576	-0.190	-0.728	-0.727	-0.446	-0.435	-0.524	-0.107	-0.843	-0.730
	(2.055)	(2.042)	(1.812)	(1.702)	(1.957)	(1.339)	(2.055)	(2.033)	(1.757)	(1.746)	(2.093)	(1.546)	(2.042)	(2.018)
nearest shopping center	-0.842	-0.849	-0.363	-0.273	-0.452	-0.121	-0.960	-0.932	-0.430	-0.421	-0.613	-0.197	-0.796	-0.750
	(2.569)	(2.557)	(2.007)	(1.809)	(2.438)	(1.636)	(2.569)	(2.564)	(2.069)	(2.047)	(2.646)	(2.011)	(2.554)	(2.518)
Census Tract Dummies							x	x	x	x	x	x		x
Maximum $\lvert \beta_k \rvert$	1.18	1.26	0.93	1.08	0.88	0.62	1.03	1.05	0.85	0.83	2.55	0.69	1.29	1.23
Average $\lvert \beta_k \rvert$	0.60	0.56	0.50	0.45	0.37	0.22	0.55	0.52	0.46	0.44	0.60	0.25	0.55	0.76
Average S_k	1.53	1.53	1.49	1.49	2.20	1.48	1.53	1.53	1.50	1.50	3.41	2.21	1.53	1.58
Average β_k	-0.49	-0.48	-0.33	-0.25	-0.35	-0.22	-0.49	-0.46	-0.30	-0.29	-0.59	-0.25	-0.47	-0.50

Table 6: Errors in Measuring MWTP When Some Housing Attributes are Unobserved, *N=200*

Mean Error / Mean True Price
(Standard Deviation of Error / Mean True Price)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Spatial Error Model	Non-parametric
bathrooms	0.223	0.115	-0.083	-0.065	0.104	0.245	-0.057	-0.198	-0.107	-0.118	-0.236	-0.146	0.038	-0.923
	(1.146)	(1.161)	(1.212)	(1.231)	(1.794)	(1.796)	(1.146)	(1.142)	(1.211)	-1.210	(1.944)	(1.583)	(1.156)	(1.166)
acreage	-0.489	-0.544	0.460	0.507	-0.197	-0.331	-0.404	-0.469	0.784	0.714	-0.007	-0.190	-0.502	-0.518
	(0.958)	(0.946)	(1.309)	(1.390)	(1.010)	(1.197)	(0.958)	(0.948)	(1.452)	-1.391	(1.089)	(1.232)	(0.947)	(0.975)
garage	-0.443	-0.244	-0.501	-0.477	-0.430	-0.473	-0.428	-0.247	-0.441	-0.429	-0.471	-0.455	-0.243	-0.866
	(1.463)	(1.436)	(1.438)	(1.438)	(1.802)	(1.585)	(1.463)	(1.435)	(1.436)	-1.437	(1.856)	(1.541)	(1.435)	(1.487)
fireplace	-0.871	-0.604	-0.741	-0.694	-0.929	-0.495	-0.878	-0.678	-0.721	-0.713	-0.935	-0.622	-0.629	-1.000
	(1.494)	(1.479)	(1.483)	(1.483)	(1.907)	(1.741)	(1.494)	(1.482)	(1.483)	-1.482	(2.465)	(1.701)	(1.480)	(1.494)
main heated living area	1.016	0.782	0.724	0.738	0.632	0.637	0.504	0.319	-0.035	-0.019	-0.034	0.123	0.699	2.004
	(1.090)	(1.194)	(1.272)	(1.316)	(1.787)	(1.629)	(1.090)	(1.131)	(1.169)	-1.168	(1.842)	(1.467)	(1.181)	(1.279)
age	-0.986	-0.982	-0.746	-0.686	-0.933	-0.324	-0.950	-0.945	-0.608	-0.619	-0.875	-0.391	-0.976	-0.904
	(1.890)	(1.888)	(1.707)	(1.658)	(1.874)	(1.438)	(1.890)	(1.884)	(1.622)	-1.633	(1.862)	(1.453)	(1.887)	(1.898)
nearest park	-0.666	-0.643	-0.195	-0.144	-0.363	-0.126	-0.670	-0.652	-0.300	-0.308	-0.450	-0.177	-0.735	-0.557
	(2.055)	(2.028)	(1.654)	(1.611)	(1.981)	(1.562)	(2.055)	(2.029)	(1.694)	-1.714	(1.968)	(1.480)	(2.034)	(2.055)
Census Tract Dummies							x	x	x	x	x	x		x
Maximum $\lvert \beta_k \rvert$	1.02	0.98	0.75	0.74	0.93	0.64	0.95	0.94	0.78	0.71	0.93	0.62	0.98	2.00
Average $\lvert \beta_k \rvert$	0.67	0.56	0.49	0.47	0.51	0.38	0.56	0.50	0.43	0.42	0.43	0.30	0.55	0.97
Average S_k	1.44	1.45	1.44	1.45	1.74	1.56	1.44	1.44	1.44	1.43	1.86	1.49	1.45	1.48
Average β_k	-0.32	-0.30	-0.15	-0.12	-0.30	-0.12	-0.41	-0.41	-0.20	-0.21	-0.43	-0.27	-0.34	-0.39

Table 7: Errors in Measuring MWTP When All Housing Attributes are Observed, *N=2000*

Mean Error / Mean True Price
(Standard Deviation of Error / Mean True Price)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
bathrooms	0.064	-0.073	-0.208	-0.099	-0.310	-0.039	0.076	-0.058	-0.066	-0.070	-0.266	-0.025
	(1.046)	(1.036)	(1.126)	(1.174)	(1.023)	(1.063)	(1.046)	(1.037)	(1.147)	-1.145	(0.986)	(1.059)
acreage	-0.483	-0.561	1.383	1.565	0.218	-0.014	-0.488	-0.570	1.302	1.272	0.130	-0.039
	(0.828)	(0.816)	(1.769)	(2.153)	(0.846)	(1.145)	(0.828)	(0.816)	(1.724)	-1.703	(0.842)	(1.139)
garage	-0.566	-0.346	-0.560	-0.497	-0.415	-0.527	-0.479	-0.282	-0.480	-0.464	-0.359	-0.532
	(1.476)	(1.446)	(1.451)	(1.448)	(1.539)	(1.515)	(1.476)	(1.445)	(1.448)	-1.448	(1.590)	(1.516)
fireplace	-0.960	-0.702	-0.822	-0.774	-0.821	-0.758	-0.925	-0.728	-0.771	-0.760	-0.896	-0.758
	(1.473)	(1.460)	(1.464)	(1.461)	(1.520)	(1.469)	(1.473)	(1.460)	(1.462)	-1.461	(1.526)	(1.468)
main heated living area	0.930	0.664	0.017	-0.012	0.195	-0.024	0.712	0.470	-0.137	-0.138	0.157	-0.022
	(0.989)	(1.074)	(1.120)	(1.151)	(1.090)	(0.982)	(0.989)	(1.045)	(1.095)	-1.093	(1.045)	(0.978)
age	-0.980	-0.976	-0.609	-0.477	-0.901	-0.412	-0.962	-0.957	-0.604	-0.605	-0.899	-0.421
	(1.894)	(1.891)	(1.620)	(1.524)	(1.878)	(1.444)	(1.894)	(1.890)	(1.617)	-1.620	(1.878)	(1.450)
median household income	0.295	0.368	-0.148	-0.094	-0.187	-0.037	0.079	0.116	-0.377	-0.359	-0.093	-0.089
	(1.078)	(1.100)	(1.183)	(1.236)	(1.129)	(1.121)	(1.078)	(1.079)	(1.147)	-1.148	(1.094)	(1.172)
median time to work	-0.116	-0.084	-0.418	-0.473	-0.046	-0.044	-0.060	0.067	-0.411	-0.396	0.060	-0.093
	(1.775)	(1.756)	(1.682)	(1.676)	(1.752)	(1.382)	(1.775)	(1.760)	(1.682)	-1.682	(1.883)	(1.427)
% under 18	-1.520	-1.602	-0.841	-0.756	0.095	-0.185	-0.936	-0.926	-0.433	-0.437	-0.509	-0.201
	(1.285)	(1.352)	(1.288)	(1.298)	(1.770)	(1.324)	(1.285)	(1.282)	(1.320)	-1.319	(1.553)	(1.320)
nearest park	-0.877	-0.854	-0.526	-0.420	-0.585	-0.124	-0.710	-0.697	-0.367	-0.366	-0.597	-0.138
	(2.035)	(2.021)	(1.729)	(1.582)	(1.909)	(1.048)	(2.035)	(2.007)	(1.637)	-1.637	(1.899)	(1.044)
nearest shopping center	-0.894	-0.907	-0.306	-0.204	-0.443	-0.045	-0.973	-0.971	-0.248	-0.250	-0.438	-0.039
	(2.518)	(2.509)	(1.828)	(1.559)	(2.345)	(1.301)	(2.518)	(2.515)	(1.779)	-1.783	(2.354)	(1.247)
Census Tract Dummies							x	x	x	x	x	x
Maximum $\lvert \beta_k \rvert$	1.52	1.60	1.38	1.56	0.90	0.76	0.97	0.97	1.30	1.27	0.90	0.76
Average $\lvert \beta_k \rvert$	0.70	0.65	0.53	0.49	0.38	0.20	0.58	0.53	0.47	0.47	0.40	0.21
Average S_k	1.49	1.50	1.48	1.48	1.53	1.25	1.49	1.49	1.46	1.46	1.51	1.26
Average β_k	-0.46	-0.46	-0.28	-0.20	-0.29	-0.20	-0.42	-0.41	-0.24	-0.23	-0.34	-0.21

Table 8: Errors in Measuring MWTP When Some Housing Attributes are Unobserved, *N=2000*

Mean Error / Mean True Price
(Standard Deviation of Error / Mean True Price)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quad.	Box-Cox Quad.	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quad.	Box-Cox Quad.	Linear	Semi-Log	Double-Log	Box-Cox Linear	Quad.	Box-Cox Quad.
bathrooms	0.43	0.30	0.03	-0.14	0.22	0.38	0.13	0.00	-0.01	-0.04	-0.17	-0.02	0.10	-0.02	-0.03	-0.07	-0.22	-0.07
	(1.05)	(1.07)	(1.16)	(1.09)	(1.04)	(1.15)	(1.05)	(1.04)	(1.16)	(1.14)	(0.99)	(1.10)	(1.05)	(1.04)	(1.15)	(1.13)	(0.99)	(1.09)
acreage	-0.52	-0.60	0.86	1.01	-0.05	-0.21	-0.49	-0.57	1.20	1.07	0.04	-0.10	-0.48	-0.57	1.24	1.10	0.07	-0.11
	(0.83)	(0.82)	(1.49)	(1.46)	(0.84)	(0.95)	(0.83)	(0.82)	(1.67)	(1.52)	(0.84)	(1.09)	(0.83)	(0.82)	(1.69)	(1.51)	(0.84)	(1.10)
garage	-0.53	-0.31	-0.57	-0.51	-0.47	-0.54	-0.47	-0.28	-0.48	-0.46	-0.45	-0.53	-0.47	-0.29	-0.49	-0.48	-0.44	-0.54
	(1.48)	(1.45)	(1.45)	(1.49)	(1.53)	(1.54)	(1.48)	(1.44)	(1.45)	(1.45)	(1.46)	(1.51)	(1.48)	(1.44)	(1.45)	(1.45)	(1.44)	(1.51)
fireplace	-0.89	-0.63	-0.77	-0.70	-0.96	-0.65	-0.92	-0.72	-0.76	-0.76	-0.91	-0.69	-0.91	-0.74	-0.77	-0.78	-0.93	-0.72
	(1.47)	(1.46)	(1.46)	(1.48)	(1.56)	(1.50)	(1.47)	(1.46)	(1.46)	(1.46)	(1.52)	(1.48)	(1.47)	(1.46)	(1.46)	(1.46)	(1.53)	(1.47)
main heated living area	1.39	1.14	0.92	0.97	0.92	0.81	0.82	0.58	0.02	0.03	0.28	0.19	0.70	0.46	-0.12	-0.11	0.17	0.05
	(0.99)	(1.16)	(1.30)	(1.17)	(1.10)	(1.16)	(0.99)	(1.06)	(1.12)	(1.11)	(1.03)	(1.05)	(0.99)	(1.04)	(1.10)	(1.08)	(1.04)	(1.03)
age	-0.99	-0.99	-0.76	-0.72	-0.95	-0.29	-0.96	-0.96	-0.62	-0.65	-0.90	-0.37	-0.96	-0.96	-0.61	-0.66	-0.90	-0.37
	(1.89)	(1.89)	(1.71)	(1.92)	(1.88)	(1.37)	(1.89)	(1.89)	(1.62)	(1.66)	(1.88)	(1.42)	(1.89)	(1.89)	(1.62)	(1.66)	(1.88)	(1.42)
nearest park	-0.69	-0.67	-0.11	-0.05	-0.31	0.03	-0.66	-0.64	-0.20	-0.22	-0.42	-0.04	-0.63	-0.63	-0.22	-0.25	-0.40	-0.08
	(2.04)	(2.01)	(1.51)	(2.21)	(1.94)	(1.22)	(2.04)	(2.00)	(1.55)	(1.60)	(1.93)	(1.15)	(2.04)	(2.00)	(1.56)	(1.63)	(1.91)	(1.13)
Census Tract Dummies							x	x	x	x	x	x						
Census Block Dummies													x	x	x	x	x	x
Maximum $\lvert \beta_k \rvert$	1.39	1.14	0.92	1.01	0.96	0.81	0.96	0.96	1.20	1.07	0.91	0.69	0.96	0.96	1.24	1.10	0.93	0.72
Average $\lvert \beta_k \rvert$	0.78	0.66	0.58	0.59	0.55	0.41	0.64	0.54	0.47	0.46	0.45	0.28	0.61	0.52	0.50	0.49	0.45	0.28
Average S_k	1.39	1.41	1.44	1.55	1.41	1.27	1.39	1.39	1.43	1.42	1.38	1.26	1.39	1.38	1.43	1.42	1.38	1.25
Average β_k	-0.26	-0.25	-0.06	-0.02	-0.23	-0.07	-0.36	-0.37	-0.12	-0.15	-0.36	-0.22	-0.38	-0.39	-0.14	-0.18	-0.38	-0.26

Figure 1: Wake County Housing Locations Relative to Census Tracts

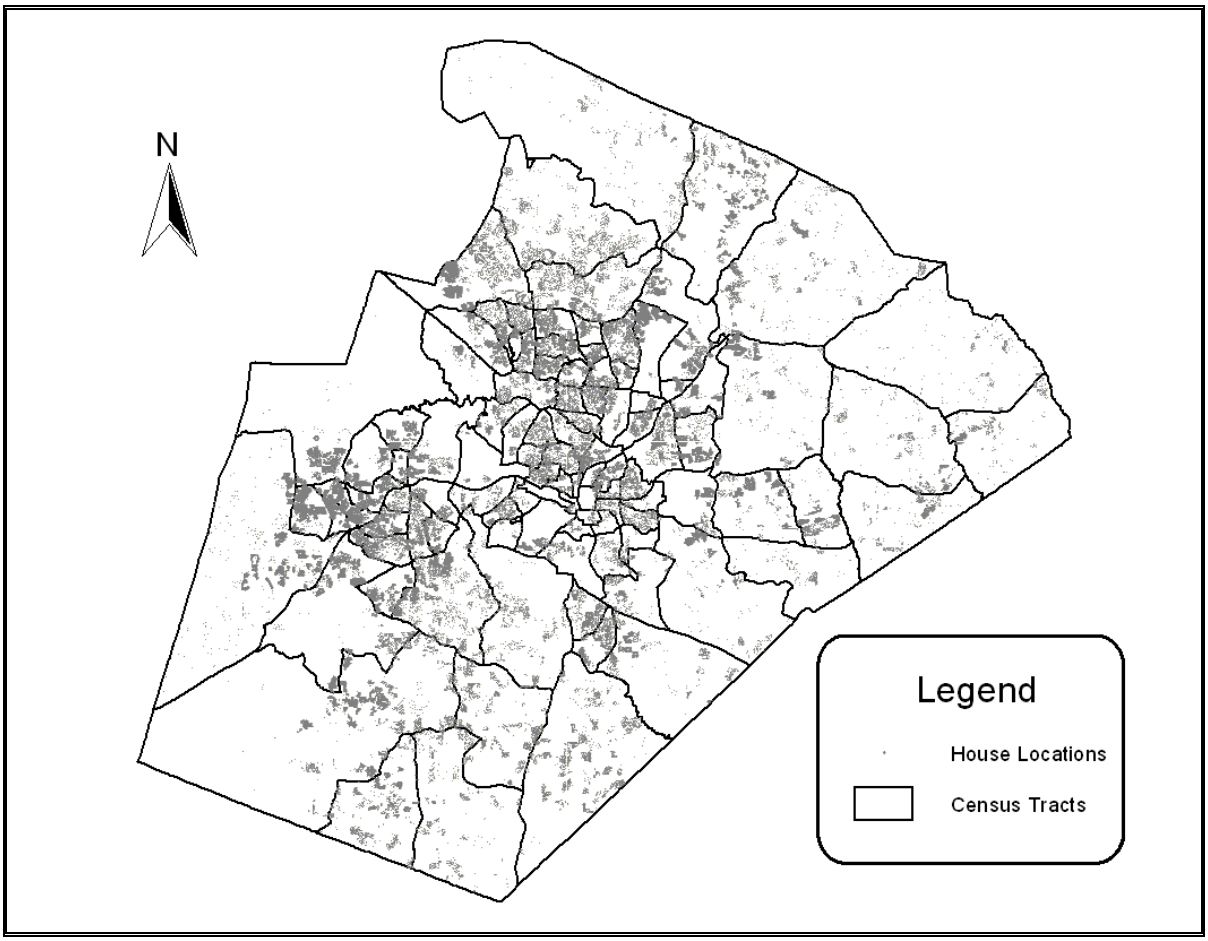
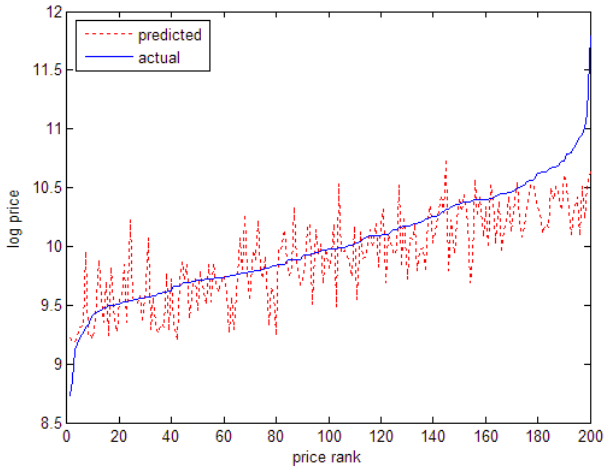
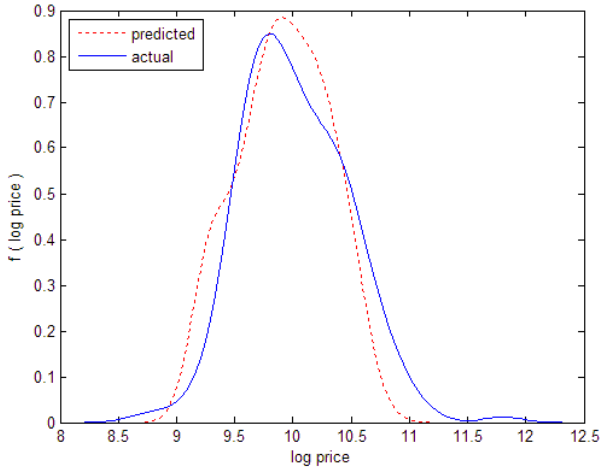


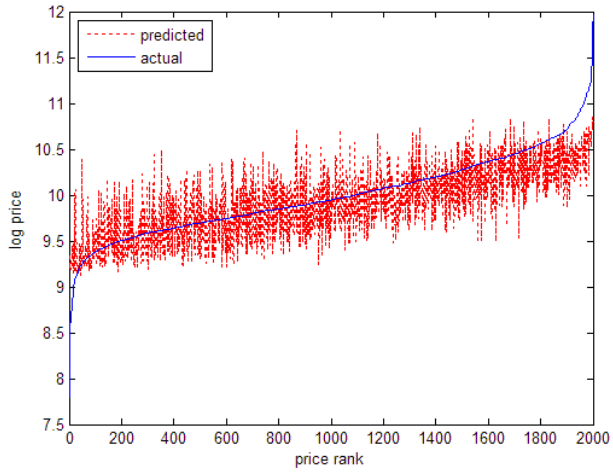
Figure 2: Reproducing the Empirical Distribution of Housing Prices in Wake County



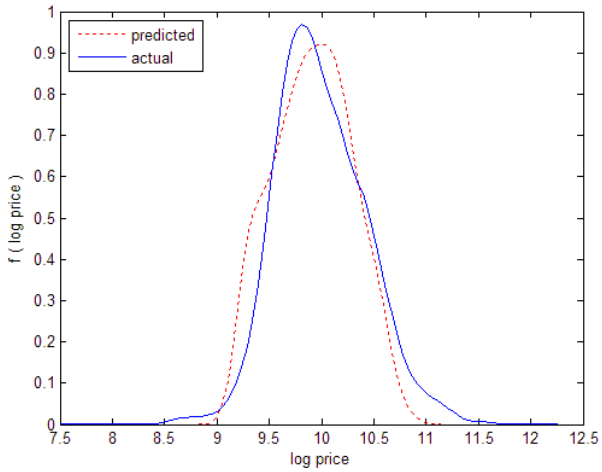
A. Deviations from price CDF (I = 200)



B. Deviations from price PDF (I = 200)



C. Deviations from price CDF (I = 2000)



D. Deviations from price PDF (I = 2000)

References

Bajari, Patrick and C. Lanier Benkard. 2005. "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach." *Journal of Political Economy*, 113(6): 1239-76.

Bajari, Patrick and Matthew E. Kahn. 2005. "Estimating Housing Demand with an Application to Explaining Racial Segregation in Cities." *Journal of Business and Economic Statistics*, 23(1): 20-33.

Banzhaf, H. Spencer. 2003. "Hedonic Pricing in Realistic Urban Structures, or What if Tiebout Called and Nobody Sorted?" *Working Paper*.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy*, 115(4): 588-638.

Bayer, Patrick and Christopher Timmins. 2007. "Estimating Equilibrium Models of Sorting across Locations." *The Economic Journal*, 117(518): 353-74.

Bin, Okmyung. 2004. "A Prediction Comparison of Housing Sales Prices by Parametric Versus Semi-Parametric Regressions." *Journal of Housing Economics*, 13(1): 68-84.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics*, 114(2): 577-99.

Chay, Kenneth Y. and Michael Greenstone. 2005. "Does Air Quality Matter? Evidence from the Housing Market." *Journal of Political Economy*, 113(2): 376-424.

Cropper, Maureen L., Leland B. Deck, and Kenenth E. McConnell. 1988. "On the Choice of Functional Form for Hedonic Price Functions." *Review of Economics and Statistics*, 70(4): 668-75.

Davis, Lucas. 2004. "The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster." *American Economic Review*, 94(5): 1693-704.

Ekeland, Ivar, James J. Heckman, and Lars Nesheim. 2004. "Identification and Estimation of Hedonic Models." *Journal of Political Economy*, 112(1): S60-S109.

Gayer, Ted, James T. Hamilton, and W. Kip Viscusi. 2000. "Private Values of Risk Tradeoffs at Superfund Sites: Housing Market Evidence on Learning about Risk." *Review of Economics and Statistics*, 82(3): 439-51.

Gençay, Ramazan and Xian Yang. 1996. "A Forecast Comparison of Residential Housing Prices by Parametric Versus Semiparametric Conditional Mean Estimators." *Economics Letters*, 52(2): 129-35.

Irwin, Elena G. 2002. "The Effects of Open Space on Residential Property Values." *Land Economics*, 78(4): 465-80.

Kim, C. , T. Phipps, and L. Anselin. 2003. "Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach." *Journal of Environmental Economics and Management, 45: 24-39*.

Kuminoff, Nicolai V. and Abdul Salam Jarrah. 2008. "Simulating Hedonic Equilibria: A Hedonic Approach." *Virginia Tech Working Paper 2008-10*.

Hall, P., Q. Li, and J. Racine. 2007. ``Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors." *Review of Economics and Statistics*, 89: 784-789.

Lancaster, Kelvin J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy*, 74(2): 132-57.

Leggett, Christopher G. and Nancy E. Bockstael. 2000. "Evidence of the Effects of Water Quality on Residential Land Prices." *Journal of Environmental Economics and Management*, 39(2): 121-44.

Li, Q., and J. Racine, 2004. Cross-Validated Local Linear Nonparametric Regression." *Statistica Sinica*, 14: 485-512.

Li, Q., and J. Racine, 2006. *Nonparametric Econometrics: Theory and Practice*. Princeton, Princeton University Press.

McCluskey, Jill J. and Gordon C. Rausser. 2003. "Stigmatized Asset Value: Is It Temporary or Long-Term?" *Review of Economics and Statistics*, 85(2): 276-85.

Palmquist, R.B. and V.K. Smith. (2002) "The use of hedonic property value techniques for policy and litigation," In: Tietenberg, T., Folmer, H. (Eds.), The International Yearbook of Environmental and Resource Economics 2002/2003. Edward Elgar, Cheltenham, UK, pp. 115-164.

Palmquist, R.B. and L.E. Danielson. (1989) "A Hedonic Study of the Effects of Erosion Control and Drainage on Farmland Values," *American Journal of Agricultural Economics*, , 55-62.

Parmeter, Christopher F. 2006. "Two-Tier Frontier and Generalized Kernel Estimation of Hedonic Price Indices." Ph.D. Dissertation, State University of New York, Binghamton (Economics), 2006.

Parmeter, Christopher F., Daniel J. Henderson, and Subal C. Kumbhakar. 2007. "Nonparametric Estimation of a Hedonic Price Function." *Journal of Applied Econometrics*, 22(3): 695-99.

Pope, Jaren C. 2008. "Buyer Information and the Hedonic: The Impact of a Seller Disclosure on the Implicit Price for Airport Noise." *Journal of Urban Economics*, 63(2): 498-516.

Poterba, James M. 1992. "Housing and Taxation: Old Questions, New Answers." *American Economic Review*, 82(2): 237-42.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, 82(1): 34-55.

Schlenker, W., W. M. Hanemann and Anthony C. Fisher. (2005) "Will U.S. Agriculture Really Benefit from Global Warming? Accounting for Irrigation in the Hedonic Approach," *The American Economic Review*, 95, 395-406.

Smith, V.K. and J.C. Huang. (1995) "Can Markets Value Air Quality? A Meta-Analysis of Hedonic Property Value Models," *Journal of Political Economy*, 103, 209-227.

Starrett, David A. 1981. "Land Value Capitalization in Local Public Finance." *Journal of Political Economy*, 89(2): 306-27.

Tinbergen, Jan. 1959. "On the Theory of Income Distribution," in *Selected Papers of Jan Tinbergen*. L.H. Klaassen, L.M. Koych and H.J. Witteveen eds. Amsterdam: North Holland.

Wheaton, William C. 1974. "Linear Programming and Locational Equilibrium: The Herbert-Stevens Model Revisited." *Journal of Urban Economics*, 1(3): 278-87.