

# **Specification of Hedonic Price Functions: Guidance for Cross-Sectional and Panel Data Applications**

By

Nicolai V. Kuminoff, Christopher F. Parmeter, and Jaren C. Pope



Working Paper No. 2009-02

January 2009

Department of Agricultural and Applied Economics  
Virginia Polytechnic Institute and State University  
Blacksburg, VA 24061

<http://www.aec.vt.edu/aec/>

# Specification of Hedonic Price Functions: Guidance for Cross-Sectional and Panel Data Applications

Nicolai V. Kuminoff

Applied Economics  
Virginia Tech  
540.231.5382  
[kuminoff@vt.edu](mailto:kuminoff@vt.edu)

Christopher F. Parmeter

Applied Economics  
Virginia Tech  
540.231.0770  
[parms@vt.edu](mailto:parms@vt.edu)

Jaren C. Pope\*

Applied Economics  
Virginia Tech  
540.231.4730  
[jcpope@vt.edu](mailto:jcpope@vt.edu)

Paper prepared for presentation at the AERE sessions of the American Economic Association

Meeting: San Francisco, California, January 3-5, 2009

---

\* We thank Dan McMillen and conference participants at the American Agricultural Economics Association Meetings in Orlando, Florida. Support from the Virginia Agricultural Experiment Station is gratefully acknowledged.

# Specification of Hedonic Price Functions: Guidance for Cross-Sectional and Panel Data Applications

**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—including panel data methods used in quasi-experimental analyses focused on removing omitted variable bias. The purpose of our paper is to update and extend the CDM (1988) simulations to investigate current issues in hedonic modeling. Three important results obtained from our theoretically consistent and carefully calibrated simulation include: (i) Using larger sample sizes than CDM and including spatial dummy variables appears to “rehabilitate” the more flexible specifications in the face of omitted variables, (ii) Exploring modern quasi-experimental and panel hedonic methods we find that in the presence of time-varying omitted variables that a first-difference specification can perform worse than OLS using pooled cross-sections, and (iii) The difference-and-difference model with spatial dummies is most accurate in identifying MWTP in situations where there is a time dimension to the data and we move between equilibria.

**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. Applied to housing markets 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 Irwin (2002) to understand the effects of open space on property values, 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 (e.g. Chay and Greenstone (2005), Davis (2004), and Linden and Rockoff (2008)).

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), henceforth CDM. 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 are estimated. This includes quasi-experimental and panel data techniques, as well as spatial econometric 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 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.

To accomplish the purpose of this paper we conduct a theoretically consistent Monte Carlo simulation. The simulation is carefully calibrated using housing data from Wake County, North Carolina that has been well documented in the literature. We use a numerical algorithm to solve for a single hedonic equilibrium to look at hedonic functional form performance in a cross-sectional setting. Later we allow some of the housing characteristics to change over time in a realistic way so that we can analyze the performance of panel data methods used in quasi-experimental analysis, relative to cross-sectional methods.

There are several important results. First, our re-analysis of the CDM functional forms

showed that when we increase the sample size and include spatial dummy variables that the flexible functional forms can once again out-perform more linear specifications in terms of estimating the MWTP of housing attributes when there are omitted variables. Furthermore, throughout our simulations (both cross-sectional and panel) the linear and semi-log models consistently the poorest performing. Second, we find that the difference-in-difference model with spatial dummies outperforms both first difference and pooled cross-sections models when we use changes of housing characteristics over time to try and identify MWTP. This is likely due to the fact that this estimator helps to account for differences in the shape of the hedonic price function before and after a large shock to a spatial amenity. Finally, we find that in the presence of time-varying omitted variables, a first-difference (or repeat-sales in the hedonic literature) specification can actually perform worse than simple ordinary-least-squares regression of pooled cross-sections.

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

## **2. 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, 2008a) 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 functional 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 CDM's (1988) Monte Carlo analysis of how the accuracy in predicting MWTP varies across competing functional form assumptions. The distinguishing feature of their study 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.<sup>1</sup>

In the 20 years that have passed since CDM's study, advances in microeconomic methods, together with the increasing availability of spatially delineated micro data, have changed the way hedonic models are estimated. 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

---

<sup>1</sup> Most model specification tests involve the 'fit' of the model. However, the conventional statistical tests of model specification do not focus directly on the gradients of the function, whereas the typical hedonic study aims to estimate marginal effects.

papers.<sup>2</sup> Table 1 compares the features of these studies to CDM.<sup>3</sup>

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 cities 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 17,000 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 spatial 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 more structure on the spatial relationship between unobserved variables and housing

---

<sup>2</sup> A complete list of these papers is provided in a supplemental appendix available from the authors upon request.

<sup>3</sup> 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.

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.

This evolution of the literature suggests that it is time to revisit the issue of functional form. The intent of our analysis is to provide information about the relative performance of the “traditional” hedonic techniques described by CDM, with more “modern” techniques in accurately estimating MWTP. More specifically, the literature motivates us to focus on the use of spatial dummy variables, spatial regression, and panel data methods such as first-difference and difference-in-difference estimation strategies.<sup>4</sup> Furthermore, the modern literature motivates us to conduct our simulations using larger sample sizes that are necessary to implement the modern functional form techniques. To the best of our knowledge, there is no existing evidence on the relative performance of these different techniques within a theoretically consistent simulation framework.

---

<sup>4</sup> We also compare our parametric results with those from a nonparametric regression framework since few studies have used fully nonparametric methods to estimate hedonic functions (see Parmeter, Henderson and Kumbhakar 2007). However, given that our interest is on estimation of the marginal effects as opposed to the overall fit of the model we feel that including them in our analysis would not provide an equal comparison given the sample sizes with which we are using in our simulation studies.

### 3. Simulation Framework

In order to investigate how our ability to accurately estimate MWTP depends on how we control for omitted variable bias and the use of panel data, we follow CDM in developing a theoretically consistent Monte Carlo simulation. After briefly reviewing the equilibrium concept and numerical algorithm we use to solve for a hedonic equilibrium, we then summarize the features of the data we use to simulate the housing 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  $j = 1, \dots, J$  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,  $\alpha$ . Let the population of households be indexed from  $i = 1, \dots, N$ . Each household is assumed to choose a specific house and a quantity of  $c$  that maximize its utility subject to a budget constraint:

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

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.<sup>5</sup> In order to define this concept more formally, let  $b_{ij}$  denote household  $i$ 's bid for the  $j^{\text{th}}$  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:

$$b_{ij} = \max_i \{ b_{ij} \} \text{ iff } A_{ij} = 1, \quad (2)$$

$$\sum_i A_{ij} = \sum_j A_{ij} = 1. \quad (3)$$

In words, each household occupies exactly one home, for which it has the maximum bid.<sup>6</sup>

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  $c$ .

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

Inserting (4) into the budget constraint and rearranging terms allows a household's maximum

---

<sup>5</sup> See Bayer and Timmins (2007) for a discussion of equilibria and estimation in location choice model with endogenously determined public goods.

<sup>6</sup> Equations (2)-(3) are equivalent to the equilibrium concept defined in equations (2)-(4) of CDM.

willingness-to-pay for a home to be expressed as a function of its characteristics and the household's income, preferences, and utility.

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

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.

Given information about preferences, income, and the stock of housing we solve for equilibrium housing prices using the Iterative Bidding Algorithm developed by Kuminoff and Jarrah (2008). The IBA iterates over a series of 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. While the IBA is based on the same equilibrium concept as the assignment algorithm used by CDM, it 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 a large number of spatial fixed effects. We use the algorithm to simulate hedonic equilibria in Wake County, North Carolina.

### **3.2 Simulating Hedonic Equilibria in Wake County's Market for Housing**

According to the 2000 census there were approximately 628,000 people living in Wake 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 276,000 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.

There are a number of reasons why Wake County provides an ideal setting for a simulation exercise aimed at understanding empirical concerns in the hedonic literature. First, there are thousands of transactions that occur in this housing market every year, giving us a large data set to work with. Second, in addition to a rich set of structural housing characteristics, the data contain information on a variety of amenities that vary within and between neighborhoods. 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. Finally, a major interstate infrastructure change that began in the late 1990's provides us with a realistic way to shock the simulated equilibrium for our analysis of panel data and quasi-experimental methods.

Simulating a hedonic equilibrium 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 several previous studies including Fulcher (2002), Pope (2008a; 2008b) and Phaneuf et al. (2008). The data spans the years 1992 to 2000 and contains approximately 104,000 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.<sup>7</sup> In our simulations, each of these 11 variables enters into the utility function. Note that this is approximately the same number of characteristics used in CDM.<sup>8</sup> The average house in the dataset sells for approximately \$201,000, has 2.5 baths, is on a .5 acre lot, does not have a garage, has a fireplace, has 1900 square feet of heated living space, is about 10 years old, is located in a census tract where median household income is \$68,000, 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

---

<sup>7</sup> 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,  $\tau$  is the owner’s marginal tax rate and is equal to 15% according to Walsh (2007),  $\tau_p$  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),  $\delta$  is depreciation set to 2% according to Poterba (1992) and  $\pi$  is land appreciation rate set to 3.19% using the average of the BLS Housing Price Index over the 1992 to 2000 time frame.

<sup>8</sup> In the simulations we use 11 characteristics for the scenarios where all housing attributes are observed whereas CDM used 12 characteristics.

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. Figure 2 shows these same census tracts in relation to a new interstate beltline that was constructed in Wake County during our study period. We use the resulting “shock” on workers commute times to analyze the empirical performance of quasi-experimental hedonic models that exploit discrete shocks as a strategy to identify MWTP.<sup>9</sup> The darker the shading of the census tract, the greater the reduction in commute times in our analysis.

We represent each household’s utility from a vector of housing characteristics,  $X_j$ , using three different parametric specifications: Cobb-Douglas, Translog, and Diewert:

$$U_{ij} = \ln(c) + \sum_j \alpha_{ij} g(X_j) + \frac{1}{2} \sum_j \sum_k \beta_{jk} h(X_j) h(X_k). \quad (7)$$

Diewert:	$g(x) = h(x) = \sqrt{x}$
Translog:	$g(x) = h(x) = \ln(x)$
Cobb-Douglas:	$g(x) = \ln(x), h(x) = 0$

In the Translog and Cobb-Douglas versions of the simulation, only the continuously varying variables were transformed. In other words, for the fireplace and garage dummy variables, we set  $g(1) = 1$  and  $g(0) = 0$ . 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

---

<sup>9</sup> See Parmeter and Pope (2008) for an extensive discussion on quasi-experiments and hedonic property value methods.

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 simulation are generated by solving for 100 hedonic equilibria. 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.<sup>10</sup> 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 3 contrasts the difference between the predicted and observed distributions of prices on a representative Monte Carlo replication. The solid line in panel A of Figure 3 represents the empirical cumulative distribution function of actual prices for 200 homes in Wake County.<sup>11</sup> 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.

---

<sup>10</sup> 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.

<sup>11</sup> 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 Monte Carlo replications to estimate various specifications for the hedonic price function. We repeated the Monte Carlo experiment using sample sizes of 200 (the sample size used by CDM) and 2000. For brevity, we only report results from the n=2000 scenario. Increasing the sample size from 200 to 2000 had three unsurprising effects on our results: (i) it increased the precision of the fit of all models, (ii) it lowered the mean bias in estimates for MWTP in all specifications, and (iii) it lowered the standard deviation of bias in estimates for MWTP in all specifications, especially for the more flexible functional forms.<sup>12</sup> Most importantly, increasing the sample size allowed us to add spatial dummy variables for census tracts without concern for the loss of degrees of freedom.

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). To evaluate the relative performance of these 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  $r$ .

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

We follow CDM by using (9) to construct summary statistics for the distribution of errors in

---

<sup>12</sup> A full set of results is available from us as a supplemental appendix. It includes all results for n=200 and n=2000 for each functional form, in each omitted variable scenario, and for each assumed utility function.

estimating MWTP for the population of households. Equation (10) defines the normalized mean ( $\beta_{kr}$ ) and standard deviation ( $S_{kr}$ ) of the errors for each attribute on a given replication.

$$\beta_{kr} = \frac{\bar{e}_{kr}}{N^{-1} \sum_i MWTP_{ikr}}, \quad S_{kr} = \frac{s_{kr}}{N^{-1} \sum_i MWTP_{ikr}}, \quad k = 1, \dots, K \quad (10)$$

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  $k$ .

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.<sup>13</sup> 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

---

<sup>13</sup> 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.

estimates for MWTP.<sup>14</sup>

In our first omitted variable scenario, the econometrician observes only one of the spatially delineated attributes in Table 1, *median time to work*. That is, *distance to the nearest shopping center* and *distance to nearest park* are omitted along with two Census block variables (*median household income, % under 18*). Without any form of correction, omitting these four variables should artificially inflate estimates for the MWTP for median time to work. Table 3 illustrates that, all else constant, moving to a home that experiences a longer commute generally means moving to a lower-income community, increasing the distance to parkland, 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 using the same utility function (Diewert) as CDM. The summary measures  $|\beta_k|$  (absolute value of mean error/bias), and  $S_k$  (standard deviation of error/bias), 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 median time to work*). For example, when all 11 characteristics are observed and a semi-log model is used in the simulation with 2000 homes, estimates of the MWTP for individual characteristics differ from the true MWTP in absolute terms by 51% on average, and the maximum difference for any characteristic is 92%. The standard deviation on the bias in estimating MWTP for the 2000 individual households in each replication is 1.55, averaged across all 100 replications.

The results in Table 4 mirror those found in CDM. Moving from left to right from

---

<sup>14</sup> Spatial dummy variables are often included in the price function with the intention of “absorbing” the price effects of unobserved neighborhood characteristics.

column (1) to column (6) (the portion of the table where all variables are observed) we see that increasing the flexibility of the functional form presents a bias-variance tradeoff. Increasing flexibility in the specification for the hedonic price function decreases the average bias in estimating MWTP, while simultaneously increasing its standard deviation. Since most hedonic studies seek to estimate average MWTP for a particular characteristic we focus on the maximum bias and the mean bias as our primary criteria for comparing the different functional forms.<sup>15</sup> Based on these two criteria, the quadratic Box-Cox model outperforms all other functional forms when there are no unobserved variables, as in CDM.

We next consider a “realistic” omitted variable scenario where we want to assess the MWTP for a particular neighborhood attribute—median time to work. In this scenario we observe structural characteristics but omit other spatially delineated amenities. This produces CDM’s main result that more complex functional forms tend to be more sensitive to omitted variables. As can be seen in columns (7) through (12) of Table 4 (the portion of the table where some variables are omitted), the Box-Cox linear, quadratic and quadratic Box-Cox models perform the worst in estimating MWTP for median time to work. Although the mean and max bias and the standard deviations increase for all specifications, they increase the most for the more flexible functional forms.<sup>16</sup> Having replicated the main results from CDM, we move on to consider new strategies for addressing omitted variables bias in cross-section and panel-data settings.

---

<sup>15</sup> The standard deviation on the bias is likely to be more important for “second-stage” hedonic studies that rely on variation in MWTP across households to identify the demand for an amenity.

<sup>16</sup> While the pattern of results in Table 4 is the same as in CDM, our quadratic and quadratic Box-Cox models do not appear to perform quite as badly as theirs. This is partly due to our use of larger sample sizes and but may also reflect advances in the numerical algorithms used to solve for the Box-Cox parameters. For example, CDM used an early version of Shazam software (ver.6) which used a convergence tolerance (0.01 or 0.001) that is high by modern standards.

## 4.2 Modern Cross-Sectional Strategies for Addressing Omitted Variables

To analyze modern cross-sectional strategies for addressing omitted variable bias, we focus on two spatially delineated variables (time-to-work and nearest park) while omitting three others. We have selected the two variables of interest in a strategic way. Time to work varies discretely across census blocks and serves as an example of a neighborhood attribute (e.g. school quality, demographics, flood zone, air quality, cancer risk, etc.) Nearest park on the other hand varies continuously and serves as an example of a spatially delineated amenity (e.g. proximity to superfund site, proximity to sex offender, proximity to beach, etc.). In Table 4 it can be seen that when all variables are observed, MWTP for time-to-work tends to be slightly overestimated whereas MWTP for distance to nearest park tends to be greatly underestimated (Table 4).

Table 5 again presents the now familiar specifications and contrasts the base results with the results from specifications that include spatial dummy variables to better address omitted variable bias. In our scenario, two of the omitted variables are neighborhood attributes (household income, % under 18) and one is a spatially delineated amenity (nearest shopping center). The top half of Table 5 shows that the omitted variables place an upward bias on MWTP for time to work and nearest park. This follows from the correlation patterns in Table 3. The upward bias on time to work is the classic omitted variable story. However, notice that the upward bias actually improves estimates for nearest park compared to Table 4 when all variables are observed. This simply reflects the pattern of spatial correlation in the data (Table 3). Max bias and mean bias are both large compared to the situation where all variables are observed in Table 4. For the quadratic Box-Cox specification, average bias and max bias more than triple.

Moving to the bottom half of Table 5 we see that adding spatial dummy variables (census tract dummies) which is common practice in the modern hedonic literature, almost completely

removes the mean bias in all specifications so that mean bias is essentially the same as in Table 4.<sup>17</sup> Max bias is also greatly decreased. Notice that the quadratic and quadratic Box-Cox specifications now perform the best (ranked by max bias and mean bias)—reversing the key result of CDM. Spatial dummy variables, which we can add due to our larger sample size of 2000 observations, appear to rehabilitate the more flexible functional forms.

A potential concern for the robustness of our results in Tables 4 and 5 (and the CDM tables) is that the omitted variable scenarios or the utility function used is somehow biased in favor of the flexible functional forms. While we think the “spatial” scenario is realistic, we would also like to look at alternative omitted variable scenarios and other choices for the utility function. Therefore, we conduct what we call a “random” omitted variable scenario where three neighborhood attributes are randomly omitted on each replication. We conduct this analysis using three different utility functions—Translog, Cobb-Douglas and the Diewert utility functions.

Table 6 illustrates that the overall pattern of results is robust to the types of variables that are omitted and the functional form of the utility function. Once again, spatial dummies consistently purge the omitted variable problem. This works despite the fact that the random scenario tends to omit structural characteristics more than spatially delineated amenities (since there are more structural characteristics to omit). It can also be seen that more flexible functional forms consistently perform better when we use spatial dummies to control for omitted variables. The quadratic Box-Cox specification has the lowest max bias for all three functional forms, and

---

<sup>17</sup> We also tried specifications using census block group dummy variables (which are smaller geographic areas than census tracts), but found that the increased spatial resolution of these indicators did not significantly improve the results.

the lowest mean bias in the Diewert and Cobb-Douglas cases.<sup>18</sup>

Table 6 also shows that the standard deviation is largest for the quadratic and quadratic Box-Cox specifications in both the Diewert and Translog scenarios. It makes sense that spatial dummies adjust for the average bias to MWTP but not deviations from that average. CDM found this same result. However, we found that as we move from  $n=200$  to  $n=2000$ , this dramatically decreases the magnitude of this effect.<sup>19</sup> The intuition for this lies in the fact that increasing the number of observations makes it easier to identify the interaction terms governing the curvature of the hedonic price function.

We also explore specifications routinely used in spatial econometrics (the spatial lag and spatial error models) to see if they outperform census block dummies in controlling for omitted variables. Based on our results in Tables 4-6, we use log transformations of all continuous variables (the double log model) and the Diewert utility function. To construct spatial weights matrices for the spatial lag and spatial error models, the Euclidean distance of each home's location to all other homes are calculated. We ensure that each house has at least one neighbor to construct the spatial weights matrices necessary for estimation of both models.

Table 7 reports the results from the spatial error and spatial lag model. The average variances are quite similar to the results including the census tract dummies, but overall we see weakened performance for the spatial error and lag models. There are several factors that could be driving this finding. First, we present results based on a single specification of the spatial weights matrix. No formal theory exists about the appropriate construction of this weights matrix and therefore a different specification may lead to materially different results. Second,

---

<sup>18</sup> Despite the result that more flexible forms tend to do better on average, it is difficult to predict which form will do best for a given amenity. Furthermore, performance can vary with the specification of utility as can be seen for the time to work and nearest park results.

<sup>19</sup> This result is available in our supplemental appendix table that provides results for all specifications in the  $n=200$  case.

the nature of the spatial structure of our housing attributes could favor the spatial dummy approach. That is, if our omitted variables vary at the census tract level then it is intuitive that a model including these census tract dummies does better relative to spatial regression model. Nonetheless, our results are suggestive that the spatial dummy approach outperforms spatial regression in some realistic scenarios and furthermore has the added feature that it removes the burden of selecting the appropriate construction of the weights matrix.

Table 7 also illustrates that the spatial error and lag models perform similarly. In the spatially omitted variable setup the spatial error model has a slightly higher maximum bias than does the spatial lag model, but appears to outperform the spatial lag model in the remaining metrics. In the random omission setup we have exactly the same results. However, even though it appears that the spatial error model is preferred, the differences in the averaged metrics between them are small relative to the variance of our measures. Thus, it is difficult to prescribe the use of one model over the other.

### **4.3. Modern Panel Data Strategies for Addressing Omitted Variables**

Current studies tend to use data that span longer time periods than when CDM conducted their analysis as discussed in section 2 and illustrated by Table 1. Furthermore, many recent applications have been based on a careful argument that we observe two different equilibria across time; before and after some shock that “treats” certain houses in a housing market (e.g. Chay and Greenstone (2005)). The assumption that the hedonic price function stays the same in these quasi-experimental hedonic analyses introduces another type of measurement error since shocks to the equilibrium may change the reduced-form parameters that describe the shape of the hedonic price function (Epple [1987]).

To provide guidance on hedonic price function specification when using pooled cross-sections of housing data or a panel of houses that sold multiple times, we shock our simulated hedonic equilibrium with changes to the variables of interest and two other spatial variables. We do this in an especially realistic way for the change in our “time to work” variable. We model the change in commute times to reflect a newly constructed interstate in Wake County that in actuality did dramatically reduce commute times for certain areas of the county (Figure 2 shows the area most affected by the newly built interstate). Table 8 shows the correlation in the changes in these variables. Comparing Table 8 with Table 3 it can be seen that for the most part, there is less correlation between the changes in variables over time than across space in the initial equilibrium. After making these changes to the spatial variables, we then solve for a new hedonic equilibrium. Not all houses change ownership between the initial equilibrium and the new equilibrium, although their resale value does change. We only use data from houses that *actually sell*, just like in a real market.

We consider three types of estimators in the quasi-experimental and panel data analysis. The first is a “pooled cross-section” estimator that simply pools the data from both equilibria. We run this estimator for all six of the basic specifications with census tract dummies, doing nothing about the temporal dimension of the data. The second estimator is a “difference-in-differences” (DID) style estimator that runs OLS on the pooled sample with the census tract dummies, but also includes a time dummy, and interactions between the time dummy and all covariates. This specification nests the basic differences-in-differences model as a special case when the amenity of interest changes over time and can be represented by a dummy variable. The third estimator is a first differences (FD) model that only uses data on houses that sold twice.

The sample size is considerably smaller in this model than for the other two estimators.<sup>20</sup>

In Table 9 we summarize results by reporting averages over 100 Monte Carlo replications for each of the three specifications of the utility function (Diewert, Translog, Cobb-Douglas) for the scenario where all variables are observed. In order to make the results comparable across specifications, we report the max and mean bias for the four variables that are included in all specifications: median time to work, nearest park, median household income, and % under 18. Since the other variables do not change, we cannot attempt to measure MWTP using the FD estimator.<sup>21</sup>

In Table 9 we can see that pooled estimation with tract dummies does especially bad on estimating the MWTP for time-to-work. This is a big departure from our cross-section model where time-to-work was estimated rather well. This bias is a reflection of the fact that the fairly large shock that was introduced changed the shape of the hedonic price function between the pre and post housing market equilibria. Simply pooling the data under the assumption that the price function is the same leads to badly biased estimates. Unlike in our pooled cross-sections scenario, the DID model recognizes that the shape of the price function may change between the two equilibria. As reported in Table 9, this lowers the bias considerably, especially for time to work. The DID model dominates pooled estimation based on average and max bias in the sense that the worst DID results are better than the best pooled estimation results. Likewise, conditional on functional form, DID gives strictly better estimates for time to work and nearest park. Moving to our FD model improves the average bias slightly. However, conditional on functional form, DID tends to do better on time to work and park. The DID model also has a

---

<sup>20</sup> Nonetheless, the number of observations is likely somewhat generous for many markets where properties do not “turn over” frequently.

<sup>21</sup> Note that the DID and FD estimators do not lend themselves to Box-Cox estimation. None of the papers we surveyed attempted Box-Cox estimation of these variables. Therefore, we leave consideration of these specifications for future work.

lower max bias for linear and semi-log models. Thus, there is no clear winner between DID and FD, although both dominate pooled estimation.

Finally, we consider the case where two of our four time-variant variables are not observed. We compare the three estimators based on the mean and standard deviation of their estimates of MWTP for time-to-work and nearest park. These results are shown in Table 10. As in Table 9, the results are averaged across 100 Monte Carlo replications for each of the three specifications for the utility function.

Not surprisingly, Table 10 shows that pooled OLS performs poorly (although the linear Box-Cox and double-log models do not do too badly on nearest park). It can also be seen that when we move to DID estimation, the bias in our estimates drops considerably. The FD estimator does somewhat better than pooled OLS on time to work, however it now performs worse than pooled OLS on nearest park. Adding tract dummies to the FD model improves estimates for time to work to be about the same as DID, but now notice how large the standard deviation is on these estimates. This is because after time-differencing the data and adding census tract dummies, there is very little within-tract variation to identify time to work, leading to a large standard deviation. The FD specification does not have the same problem with nearest park since it varies continuously. However, notice that the results are slightly worse than the DID model.

Based on the results from Tables 9 and 10, the DID specification seems best suited to identify MWTP in situations where there is a long time dimension to the data and we move between equilibria. It is theoretically consistent in the sense that it recognizes that marginal implicit prices of characteristics may change in the new equilibrium. It also seems fairly robust to the presence of omitted variables.

## 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. To date, there has been no effort to update CDM by analyzing more “modern” hedonic property value methods and evaluate the relative performance of these techniques in a controlled simulation. The purpose of our paper was to fill this gap in the literature.

Our re-investigation of the “traditional” hedonic estimation through a theoretically consistent Monte Carlo simulation highlighted two important results. First we found that larger sample sizes combined with spatial dummy variables rehabilitate flexible functional forms such as the quadratic Box-Cox model. This result is consistent across a realistic “spatial” omitted variable scenario and a “random” omitted variable scenario. Second we found that while flexible forms do best on average, performance on measuring the MWTP for a particular amenity can vary with the spatial landscape, the pattern of omitted variables, and consumer preferences.

Our investigation into “modern” quasi-experimental and panel approaches to hedonic estimation provided three additional insights. First we found that in the presence of time-varying omitted variables, a first-difference specification can perform worse than naively pooling cross-

section data without regard for time. Second we found that spatial dummies can help to purge omitted variable bias in a FD model. However, the addition of spatial dummy variables can substantially decrease the degree of variation in amenities that vary discretely across neighborhoods, leading to large standard deviations on the sampling distribution for MWTP. Third we found that overall, the DID model which included spatial dummy variables and interactions between the time dummy and all the arguments of the hedonic price function seemed to perform best in the panel setting. This estimator offers a middle ground between structural and experimental approaches to hedonic modeling. It takes a quasi-experimental approach to identification while acknowledging one of the key implications of hedonic theory—that a large shock to an urban amenity may change the shape of the hedonic price function.

Hedonic property value methods have a long history in applied work within many fields of economics. Given the advent of GIS technologies and the increasing availability of digitally archived housing data that is made available by county assessors throughout the country, the potential to study interesting environmental and urban questions has grown tremendously. Furthermore, this new wave of housing data provides a temporal resolution that also makes it amenable to quasi-experimental and panel data analyses. For these reasons, we expect hedonic property value models will see increased application. We think that the results presented in this paper can provide valuable guidance for empirical specification of hedonic price functions and can be used to guide future hedonic analyses.

## References

- Anselin, Luc. 1988. *Spatial Econometrics: methods and models*, Dordrecht: Kluwer Academic.
- Bayer, Patrick and Christopher Timmins. 2007. "Estimating Equilibrium Models of Sorting across Locations." *The Economic Journal*, 117(518): 353-74.
- 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 Kenneth 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.
- Epple, Dennis. 1987. "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products." *Journal of Political Economy*, 95(1): 59-80.
- 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.
- Irwin, Elena G. 2002. "The Effects of Open Space on Residential Property Values." *Land Economics*, 78(4): 465-80.
- Kuminoff, Nicolai V. and Abdul Salam Jarrah. 2008. "Simulating Hedonic Equilibria: A Hedonic Approach." *Virginia Tech Working Paper 2008-10*.
- 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.
- 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.
- Parmeter, Christopher F., Daniel J. Henderson and Subal C. Kumbhakar. 2007. "Nonparametric Estimation of a Hedonic Price Function," *Journal of Applied Econometrics* 22(3), 695-699.
- Parmeter, Christopher F. and Jaren C. Pope. 2008. "Quasi-Experiments and Hedonic Property Value Methods," Working paper available at SSRN: <http://ssrn.com/abstract=1283705>.
- Phaneuf, Daniel J., V. Kerry Smith, Raymond B. Palmquist and Jaren C. Pope. 2008. "Integrating Property Value and Local Recreation Models to Value Ecosystem Services in Urban Watersheds," *Land Economics* 84(3), 551-572.
- Pope, Jaren C. 2008a. "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.
- Pope, Jaren C. 2008b. "Do Seller Disclosures Affect Property Values? Buyer Information and the Hedonic Model." *Land Economics*, 84(4): 551-572.
- 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.
- 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.

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

<b>Total # studies</b>		CDM	SSCI (61)
<b>Functional form</b>	# using lin-lin, log-lin, log-log		47
	# using Box-Cox		12
<b>Dummy Variables</b>	Mean share of covariates which are 0/1	17%	36%
<b>Sample Size</b>	<u>Median # observations</u>	200	1,679
	published in 1989-1998		593
	published in 1999-2008		2,066
	<u>Distribution of studies by # observations</u>		
	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
<b>Housing Market</b>	<u>Geography (#)</u>		
	smaller than a city		5
	city or county	1	24
	multiple cities or counties		28
	nation		4
	<u>Time Period (#)</u>		
	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	
<b>Space and Time</b>	# with spatial error or spatial lag structure		7
	# exploiting panel structure of data		15
	<u># with time dummies or time trend</u>		23
	day		3
	month		3
	quarter		2
	year		15
	<u># with spatial dummies</u>		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: Comparison of Basic Results to CDM**  
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	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
bathrooms	0.051 (1.406)	-0.120 (1.412)	-0.185 (1.404)	-0.190 (1.417)	-0.408 (1.553)	-0.236 (1.527)	0.825 (1.406)	0.573 (1.472)	0.674 (1.459)	0.651 (1.558)	0.500 (1.666)	0.607 (1.777)
acreage	-0.204 (1.530)	-0.313 (1.528)	2.020 (2.180)	1.788 (2.122)	0.260 (1.517)	-0.273 (1.817)	-0.225 (1.530)	-0.354 (1.527)	1.183 (1.823)	0.861 (1.715)	0.101 (1.543)	-1.070 (2.730)
garage	-0.201 (1.450)	0.352 (1.464)	0.072 (1.446)	0.334 (1.544)	-0.002 (1.520)	0.068 (1.530)	0.539 (1.450)	1.050 (1.540)	0.809 (1.509)	1.194 (1.824)	0.605 (1.528)	0.930 (1.756)
fireplace	-1.197 (1.432)	-0.557 (1.418)	-0.937 (1.430)	-0.577 (1.424)	-0.537 (1.541)	-0.367 (1.474)	-0.696 (1.432)	-0.081 (1.424)	-0.412 (1.418)	0.096 (1.500)	-0.203 (1.685)	0.278 (1.961)
main heated living area	0.639 (1.407)	0.330 (1.455)	0.107 (1.412)	-0.016 (1.423)	-0.064 (1.526)	-0.147 (1.567)	2.184 (1.407)	1.698 (1.670)	1.689 (1.551)	1.550 (1.735)	1.734 (1.615)	1.766 (1.990)
age	-0.948 (1.965)	-0.924 (1.960)	-0.130 (1.841)	-0.204 (1.865)	-0.757 (1.939)	-0.068 (1.865)	-0.921 (1.965)	-0.902 (1.958)	-0.314 (1.825)	-0.474 (1.853)	-0.806 (1.944)	-0.206 (1.897)
median household income	-0.318 (1.221)	-0.406 (1.214)	-0.488 (1.201)	-0.477 (1.200)	-0.476 (1.287)	-0.443 (1.293)	This variable has been omitted					
median time to work	0.260 (1.808)	0.269 (1.811)	0.271 (1.816)	0.374 (1.900)	0.306 (2.016)	0.252 (2.048)	1.003 (1.808)	1.123 (1.884)	1.397 (1.988)	1.761 (2.372)	1.430 (1.875)	1.622 (2.399)
% under 18	-0.790 (1.312)	-0.803 (1.303)	-0.651 (1.288)	-0.676 (1.291)	-0.255 (1.415)	-0.384 (1.351)	This variable has been omitted					
nearest park	-0.939 (1.718)	-0.834 (1.709)	-0.337 (1.658)	-0.352 (1.675)	-0.477 (1.709)	0.046 (2.002)	This variable has been omitted					
nearest shopping center	-0.700 (1.810)	-0.681 (1.795)	0.045 (1.837)	0.064 (1.858)	-0.097 (1.778)	0.263 (2.398)	This variable has been omitted					
Maximum $ \beta_k $	1.20	0.92	2.02	1.79	0.76	0.44	2.18	1.70	1.69	1.76	1.73	1.77
Average $ \beta_k $	0.57	0.51	0.48	0.46	0.33	0.23	0.91	0.83	0.93	0.94	0.77	0.93
Average $S_k$	1.55	1.55	1.59	1.61	1.62	1.72	1.57	1.64	1.65	1.79	1.69	2.07

**Table 5: “Modern” Cross-Sectional Specification Results.**

Metric for Comparison	<i>functional form</i>					
	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
<i>Census Tract Dummy Variables Excluded</i>						
$\beta_{\text{time-to-work}}$	0.61	0.64	0.90	1.16	0.96	1.18
$\beta_{\text{nearest park}}$	-0.61	-0.52	0.11	0.12	-0.20	0.40
Max $ \beta_j $	2.14	1.65	1.81	1.39	1.52	1.54
Average $ \beta_j $	0.81	0.71	0.78	0.76	0.64	0.71
Average $ S_j $	1.59	1.64	1.68	1.78	1.70	2.24
<i>Census Tract Dummy Variables Included</i>						
$\beta_{\text{time-to-work}}$	-0.08	0.08	0.06	0.22	0.09	0.21
$\beta_{\text{nearest park}}$	-0.41	-0.29	0.27	0.34	-0.19	0.35
Max $ \beta_j $	0.88	0.86	1.68	1.32	0.75	0.65
Average $ \beta_j $	0.43	0.38	0.40	0.41	0.28	0.26
Average $ S_j $	1.59	1.60	1.64	1.67	1.65	1.87

\* Three neighborhood attributes are omitted on every replication: *median household income*, *% under 18*, and *nearest shopping center*.

**Table 6: Robustness of Results to Omitted Variable Selection and Utility Functional Form**

Metric for Comparison	<i>functional form</i>					
	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
<i>Diewert Utility Function</i>						
$\beta_{\text{time-to-work}}$	-0.09	0.05	0.12	0.14	0.29	0.34
$\beta_{\text{nearest park}}$	-0.62	-0.50	-0.04	-0.14	-0.40	0.08
Max $ \beta_j $	0.88	0.86	1.52	0.90	0.75	0.64
Average $ \beta_j $	0.50	0.46	0.42	0.38	0.35	0.31
Average $ S_j $	1.55	1.56	1.58	1.58	1.61	1.74
<i>Translog Utility Function</i>						
$\beta_{\text{time-to-work}}$	0.25	0.37	0.02	0.07	0.40	0.39
$\beta_{\text{nearest park}}$	-0.59	-0.57	-0.14	-0.08	-0.36	0.19
Max $ \beta_j $	0.94	0.93	0.70	0.69	0.86	0.69
Average $ \beta_j $	0.56	0.52	0.25	0.23	0.42	0.31
Average $ S_j $	1.70	1.71	1.64	1.74	1.77	1.99
<i>Cobb-Douglas Utility Function</i>						
$\beta_{\text{time-to-work}}$	0.03	0.13	-0.14	-0.09	0.18	0.18
$\beta_{\text{nearest park}}$	-0.70	-0.69	-0.37	-0.33	-0.54	-0.15
Max $ \beta_j $	1.15	0.96	1.08	0.88	0.91	0.74
Average $ \beta_j $	0.66	0.61	0.47	0.42	0.49	0.35
Average $ S_j $	1.49	1.50	1.48	1.49	1.53	1.39

\* Three *randomly chosen* neighborhood attributes are omitted on each replication.

**Table 7: Spatial Error and Spatial Lag Model Results**

Omitted Variable Scenario	Metric for Comparison	<i>Model</i>	
		Spatial Error	Spatial Lag
Spatial	$\beta_{\text{time-to-work}}$	0.74	0.89
	$\beta_{\text{nearest park}}$	-0.17	-0.26
	Max $ \beta_j $	2.34	2.28
	Average $ \beta_j $	0.68	0.71
	Average $ S_j $	1.67	1.67
Random	$\beta_{\text{time-to-work}}$	-0.47	-0.54
	$\beta_{\text{nearest park}}$	-0.57	-0.60
	Max $ \beta_j $	1.23	1.21
	Average $ \beta_j $	0.52	0.54
	Average $ S_j $	1.62	1.62

\* The *Spatial* scenario omits three neighborhood attributes on every replication: *household income*, *% under 18*, and *nearest shopping center*.

\*\* The *Random* scenario randomly selects three variables to omit on each replication

**Table 8: Correlation Coefficients for Changes in Variables**

Spatial variation	Variable	$\Delta$ median	$\Delta$ median	$\Delta$ %	$\Delta$ nearest
Census block	$\Delta$ median household income	1.00			
Census block	$\Delta$ median time to work	0.39	1.00		
Census block	$\Delta$ % under 18	0.36	0.12	1.00	
home	$\Delta$ nearest park	0.14	0.11	0.42	1.00

**Table 9: Panel Data Estimator Results when all Variables are Observed**

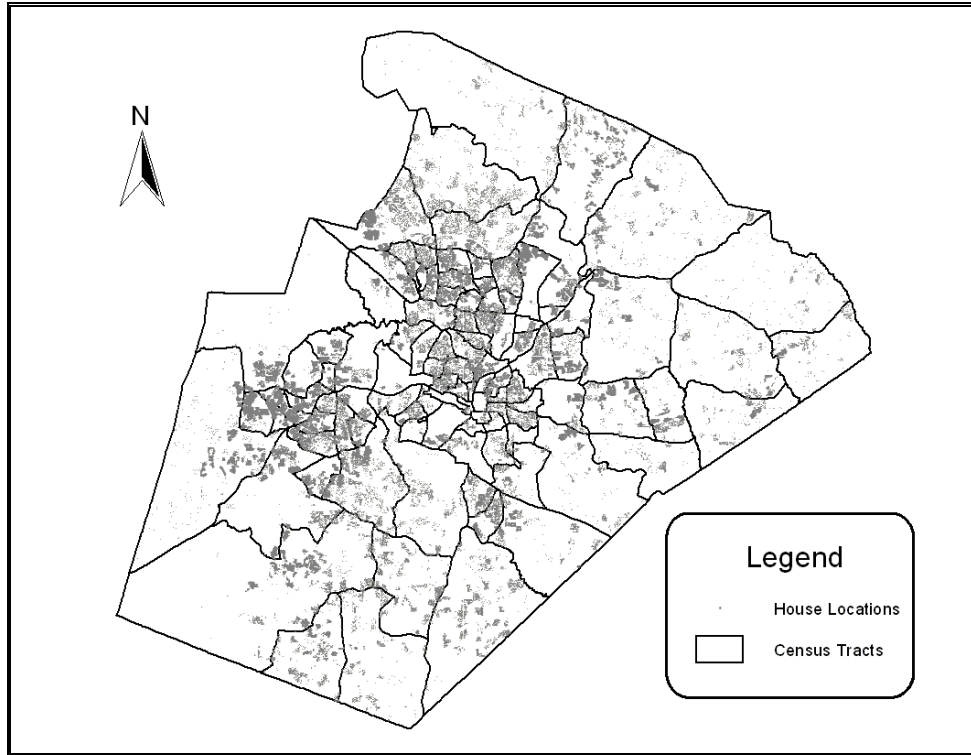
	Pooled Estimation						Difference-in-Difference			First Difference		
	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quad.	Box-Cox Quad.	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log
$ \beta_{\text{time-to-work}} $	0.82	0.76	0.87	0.89	0.91	0.96	0.19	0.11	0.39	0.33	0.34	0.15
$ \beta_{\text{nearest park}} $	0.69	0.68	0.21	0.15	0.47	0.14	0.59	0.59	0.18	0.61	0.60	0.19
Max $ \beta_j $	0.82	0.77	0.87	0.89	1.06	0.97	0.64	0.64	0.58	0.67	0.70	0.46
Average $ \beta_j $	0.64	0.64	0.60	0.61	0.66	0.59	0.40	0.40	0.38	0.35	0.38	0.27
Average $ S_j $	1.71	1.70	1.79	2.13	1.74	2.84	1.72	1.71	1.74	1.76	1.75	1.95
tract dummies	x	x	x	x	x	x	x	x	x			
average N	3643	3643	3643	3643	3643	3643	3643	3643	3643	1643	1643	1643

**Table 10: Panel Data Estimator Performance w/Omitted Neighborhood Attributes**

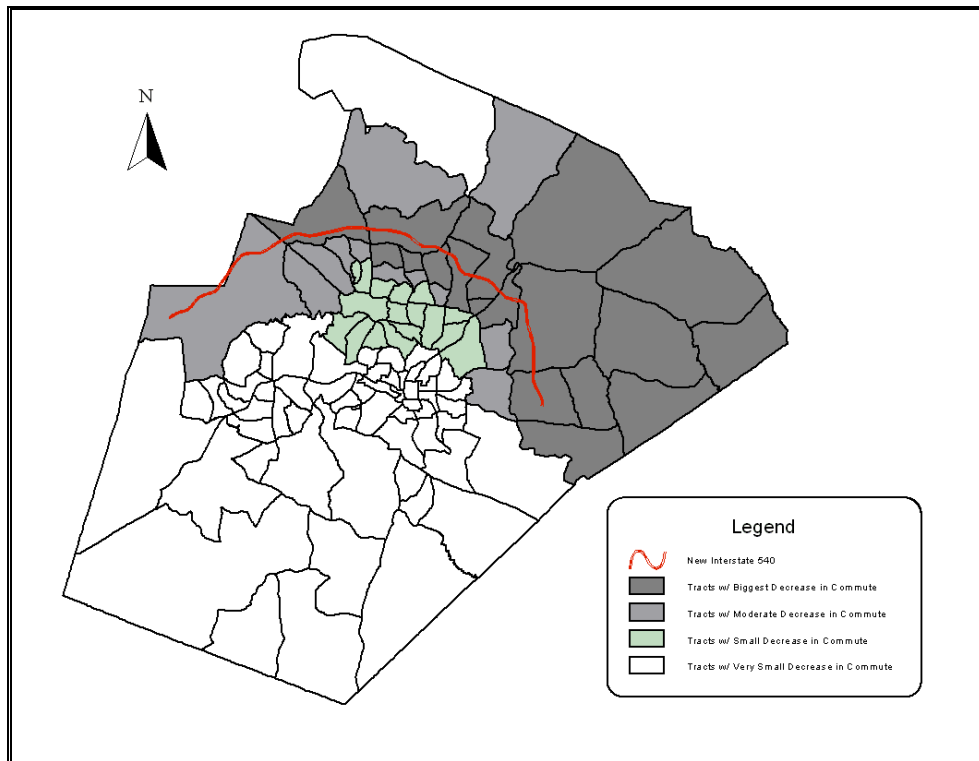
	Pooled Estimation						Difference-in-Difference			First Difference					
	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quad.	Box-Cox Quad.	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log
$ \beta_{\text{time-to-work}} $	0.69 (0.07)	0.66 (0.07)	0.81 (0.05)	0.81 (0.05)	0.67 (0.06)	0.80 (0.05)	0.41 (0.09)	0.33 (0.09)	0.43 (0.07)	0.56 (0.05)	0.52 (0.05)	0.49 (0.05)	0.44 (0.35)	0.58 (0.29)	0.38 (1.35)
$ \beta_{\text{nearest park}} $	0.62 (0.04)	0.60 (0.04)	0.18 (0.05)	0.17 (0.06)	0.25 (0.05)	0.20 (0.06)	0.54 (0.04)	0.52 (0.04)	0.15 (0.06)	0.84 (0.03)	0.81 (0.03)	0.31 (0.06)	0.73 (0.05)	0.76 (0.04)	0.18 (0.07)
Average $ \beta_j $	0.66	0.63	0.49	0.49	0.46	0.50	0.47	0.43	0.29	0.70	0.67	0.40	0.58	0.67	0.28
tract dummies	x	x	x	x	x	x	x	x	x				x	x	x
average N	3643	3643	3643	3643	3643	3643	3643	3643	3643	1643	1643	1643	1643	1643	1643

\* Three neighborhood attributes are omitted on every replication: *median household income*, *% under 18*, and *nearest shopping center*.

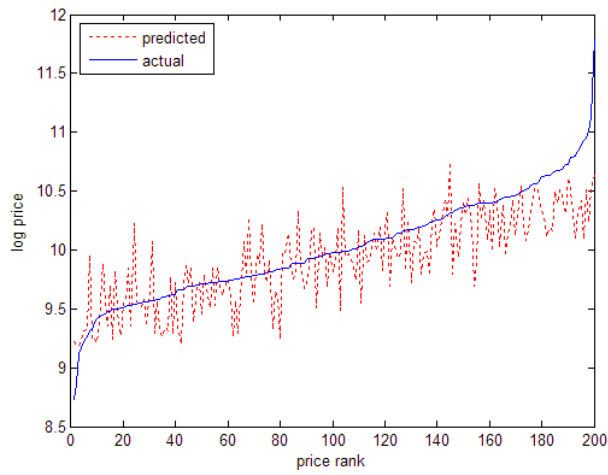
**Figure 1: Wake County Housing Locations Relative to Census Tracts**



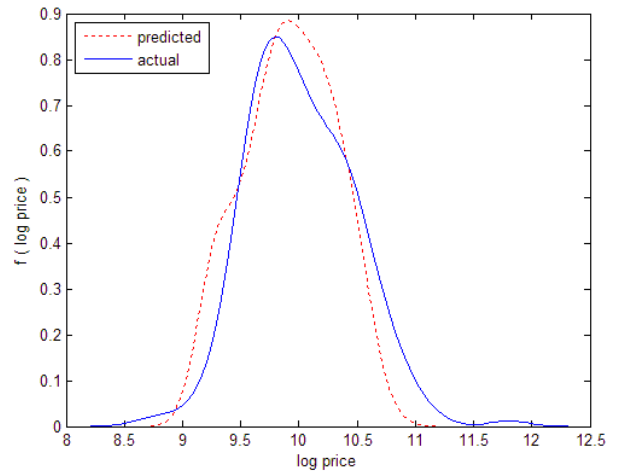
**Figure 2: Location of Interstate Project Relative to Census Tracts**



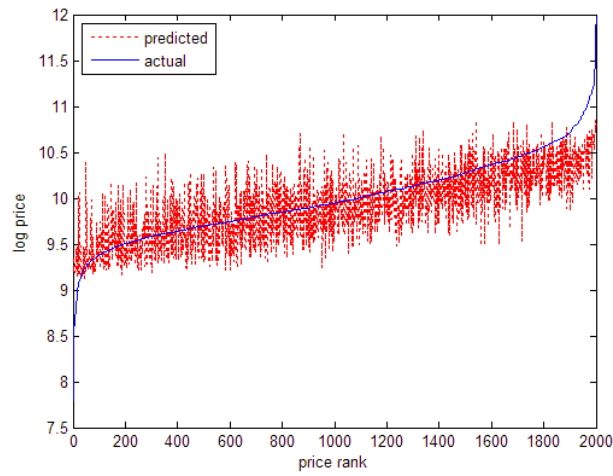
**Figure 3: Reproducing the Empirical Distribution of Housing Prices in Wake County**



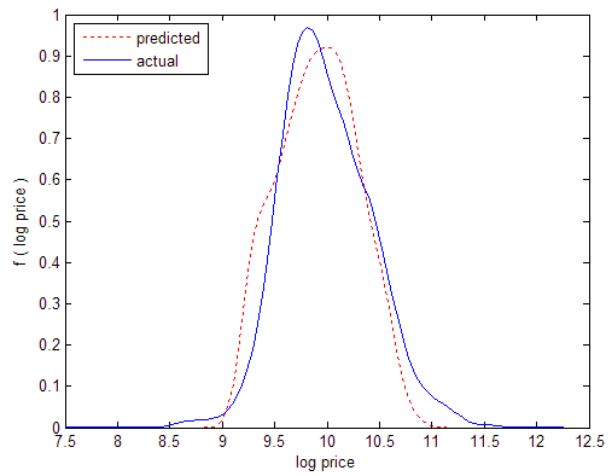
*A. Deviations from price CDF ( I = 200 )*



*B. Deviations from price PDF ( I = 200 )*



*C. Deviations from price CDF ( I = 2000 )*



*D. Deviations from price PDF ( I = 2000 )*