Averting Hidden Danger: Developing Hedonic Specifications that Mitigate Omitted Variable Bias

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

March 14, 2009

* We thank Matt Kahn, Dan McMillen, Dan Phaneuf, V. Kerry Smith, and Laura Taylor for helpful comments and suggestions on this research. We also thank seminar participants at the 2009 AEA meetings. Financial support from the Virginia Agricultural Experiment Station is gratefully acknowledged.

Averting Hidden Danger: Developing Hedonic Specifications that Mitigate Omitted Variable Bias

ABSTRACT: The hedonic method applied to housing markets is widely used to estimate the willingness-to-pay for public goods. However, there is increasing concern that omitted variables seriously undermine the method's ability to accurately identify economic values. This paper develops a theoretically consistent Monte Carlo analysis of the efficacy of modern hedonic estimators—those employing spatial fixed effects, panel data, and quasi-experimental methods—in addressing omitted variable bias and recovering nonmarket values. Our simulation results provide the first hard evidence that these new methods work. The evidence also suggests it is time to rethink some of the stylized facts about hedonic estimation.

KEY WORDS: Hedonic, functional form, omitted variable, quasi-experiment, property values

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

1. Introduction

In his seminal 1974 paper, Sherwin Rosen demonstrated 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 marginal changes in individual attributes of a differentiated product (MWTP). This result is especially important for policy evaluation because it provides the means to estimate welfare measures for changes in public goods that are not explicitly traded in formal markets, but are conveyed through the purchase of a home. Hedonic property value models are frequently used to estimate the MWTP for spatially delineated public goods 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), crime rates (Linden and Rockoff, 2008) and airport noise (Pope, 2008) to name only a few of the recent applications. Accurate estimates of MWTP for these and other nonmarket goods and services are of paramount importance to policymakers and researchers.

Over the years there has been increasing concern that misspecification of the hedonic price function can seriously undermine its ability to accurately estimate MWTP. The foremost concern is omitted variable bias. Neighborhood characteristics that matter to households, but are not observed by the econometrician, may be correlated with the amenity of interest or other independent variables. While the omitted variable problem is quite general, it is challenging to evaluate its implications for hedonic estimates of nonmarket values because housing prices and consumer welfare are simultaneously determined as equilibrium outcomes of the market clearing process (Rosen, 1974). Cropper, Deck and McConnell (1988) were the first to meet this challenge by developing a theoretically consistent framework for simulating hedonic equilibria. One of their key findings was that flexible econometric specifications for the equilibrium price function performed best when all variables were included in the model, but simpler functional forms (i.e. linear, log-linear, and log-log) performed best in the presence of omitted variables. This result was widely applied in the subsequent empirical literature. The vast majority of hedonic studies published over the last two decades have used linear, log-linear, or log-log models in an effort to hedge against the risk of omitted variable bias.

In the 20 years since Cropper, Deck, and McConnell's innovative study (henceforth CDM) econometric methodology in the hedonic literature has also evolved to directly address omitted variables. Larger sample sizes now permit the use of spatial fixed effects and the increasing availability of pooled cross-section and panel data has allowed researchers to exploit quasi-experiments to purge time-constant omitted variables (e.g. Davis, 2004; Greenstone and Gallagher, 2008; Linden and Rockoff, 2008). Surprisingly there has been no systematic effort to evaluate the relative performance of these modern techniques in dealing with omitted variables and in recovering nonmarket values. We seek to fill this void.

This paper describes a theoretically consistent Monte Carlo evaluation of the efficacy of modern hedonic methods in recovering nonmarket values when key variables are omitted from the model. Five features of our analysis go beyond CDM and allow us to inform current hedonic practice: (i) we use data on 104,000 housing sales to reproduce actual market conditions in Wake County, NC in the late 1990s, (ii) the simulated urban landscape reflects real patterns of spatial correlation between housing characteristics and multiple urban amenities, (iii) the sample size for econometric analysis is consistent with modern studies and allows the use of spatial fixed effects,

(iv) we focus on omitted urban amenities and test the robustness of our results by randomly selecting variables to omit, and (v) we simulate housing market dynamics. For realism, the hedonic dynamics are based on an actual shock to commuting patterns that occurred in Wake County during our study period. This allows us to simulate a quasi-experiment to evaluate pooled cross-section, first difference, and difference-in-difference approaches to estimation.

Overall, we ran 54,000 regressions to test 540 different models, generating more than half a billion estimates of household MWTP. While the diversity of our econometric specifications addresses Leamer's (1983) concern about the dangers of implicit model selection, the sheer scale of our analysis makes it challenging to synthesize the results. We approach this problem by first describing key results for cross-section and panel data estimators and then conducting a meta-analysis that relates model performance to modeling choices, allowing for comparison across all specifications (Banzhaf and Smith, 2007).

Our analysis yields several important results. First, adding spatial fixed effects for Census tracts to the estimated price function substantially reduces the bias from omitted variables in cross-section data. This reduction is much larger than the reduction that comes from using simpler functional forms. Moreover, the reduction is sufficiently large for "flexible" Box-Cox specifications that they outperform the linear, log-linear, and log-log models which currently dominate empirical practice.

Second, difference-in-difference and first differenced estimation further reduces the bias from *time-constant* omitted variables, beyond the gains from spatial fixed effects. These estimators can also be effective in recovering nonmarket values in the presence of *time-varying* omitted variables, if used in conjunction with spatial fixed effects.

Third, we find that housing market dynamics expand the scope for econometric misspecification. Our simulations characterize the movement between equilibria following a moderate shock to commute time and small shocks to other variables. Simply pooling the data from the two equilibria leads to a bias in estimated MWTP that is similar in magnitude to the bias from omitted variables. This bias can be averted by using a generalized difference-in-differences estimator to model implicit price dynamics.

Finally, our meta-analysis of model implementation reveals several other interesting findings, including: (i) the accuracy in estimating MWTP for neighborhood amenities is closely tied to sample size, (ii) spatial fixed effects have greater success in averting omitted variable bias than spatial error models and spatial lag models, and (iii) in a panel data setting with time-varying omitted variables, linear specifications for the first-differenced price function outperform nonlinear specifications, providing an analog to the seminal result from CDM's cross-section simulation.

The remainder of the paper proceeds as follows. Section 2 begins by briefly reviewing the evolution of the empirical hedonic literature on estimating the MWTP for public goods and environmental amenities. Section 3 describes how we designed features of the simulation to be representative of modern studies. Then section 4 reports and interprets the results, which are followed by concluding comments in section 5.

2. Empirical Specification and Omitted Variables in the Hedonic Literature

In Rosen's (1974) hedonic 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 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 property value studies treat linearity as a maintained assumption.[1] 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.[2] 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 models produce the lowest mean percentage error in estimating MWTP. This result changes when one of the variables is omitted 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" simulation have guided the subsequent empirical literature due to widespread concern about the potential for omitted variables to bias

[1] Due to space constraints, we do not work through the standard analytics of the hedonic property value model. Readers interested in an analytical review of the literature are directed to Palmquist (2005).
[2] We know of only two other studies that have developed similar theoretically consistent simulations. Cropper et al. (1993) compare hedonic and random-utility models, and an unpublished manuscript by Banzhaf (2003) explores the consequences of discreteness in the density of housing characteristics.

hedonic estimates for MWTP. It has become standard practice to adopt one of the more parsimonious functional forms from CDM's study to represent the equilibrium hedonic price function with little or no specification testing. This is somewhat surprising because modern hedonic studies use data and econometrics which have little in common with the simulations designed by CDM.

In the 20 years that have passed since CDM's study, advances in microeconometric methods, together with the increasing availability of micro data, have changed the way hedonic models are estimated. To document these changes, we reviewed the 123 studies published between November 1988 and January 2009 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, ecolabeled paper towels, appliances, collectable coins, television, fish, forestry, and agricultural land. Narrowing the focus to residential property value studies decreased the size of our sample to 69 papers published in 35 different journals. This is clearly not a random sample and one might be concerned that it would overstate CDM's influence on the literature. Therefore, we also reviewed every property value hedonic study published in the *American Economic Review*, *Journal of Political Economy*, *Review of Economics and Statistics*, and the *Quarterly Journal of Economics* during the same 20-year period. We abbreviate this collection of papers as ECON. Only 5 of the 85 papers we reviewed show up in both samples. Notably 60% of the ECON papers were published in the last 5 years and 90% were published in the last 10 years, underscoring the growing general interest in this line of research.[3]

[3] A complete list of the SSCI and ECON papers will be provided by the authors upon request. The ECON papers were drawn from our archives and a supplemental search on Google Scholar that returned every paper published in the four journals that included any one of the following five phrases anywhere in the article: *hedonic*, *property values*, *housing prices*, *repeat sales*, and *capitalization*.

Table 1 compares the features of CDM to the papers in our SSCI and ECON samples.[4] The influence of CDM on the choice of functional form is immediately apparent. 80% of the SSCI studies and 94% of the ECON studies 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 the SSCI studies published during the past ten years (2,459) more than quadrupled from the previous ten year period (593) which was nearly triple the number of observations used by CDM (200). Median sample size is even larger in the ECON studies published during the past ten years (13,414). 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, 51% of the SSCI papers use data from multiple cites or counties and 75% use sales data over more than two years. These percentages are even larger for the ECON sample where 75% of the studies use data from multiple cites or counties and 87% use more than two years of data.

Over the past 20 years, the literature has also evolved to address omitted variables directly. More than half the studies in the SSCI sample and three quarters of the studies in the ECON sample use spatial fixed effects to absorb the effect of unobserved amenities that vary between cities or between "neighborhoods" within a city (such as census tracts or school districts). A smaller share of SSCI papers (14%) use spatial econometrics to impose more

[4] 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.

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 conduct quasi-experimental analyses using property value data. Most ECON studies have a quasi-experimental design and although quasi-experimental papers are a minority in the full SSCI sample, most have been published in the past few years.[5] 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 it is time to revisit the issue of empirical hedonic specification. The principal difficulty in undertaking this type of analysis is that the set of possible specifications is infinite. Therefore, we investigate the correspondence between popular regions in assumption space and regions in the inference space (Leamer, 1983). Our main objective is to compare the performance of the "traditional" hedonic techniques investigated by CDM with the performance of the "quasi-experimental" techniques that have dominated empirical practice in the recent hedonic literature. Specifically, we investigate how the accuracy in estimating MWTP varies with the use of spatial fixed effects, spatial regression techniques, and panel data strategies for addressing the confounding influence of omitted variables. Our analysis provides the first evidence on the relative performance of these different techniques within a theoretically consistent simulation framework.

[5] We define a hedonic study as being quasi-experimental if there is an exogenous discontinuity (in time or space) that generates clear treatment group(s) and clear control group(s) which are used to estimate treatment effect(s). Many older studies satisfy these criteria without using the modern quasi-experimental terminology.

3. Simulation Framework

In order to investigate how our ability to accurately estimate MWTP depends on how we control for omitted variables, we follow CDM in developing a theoretically consistent Monte Carlo simulation. After briefly reviewing the equilibrium concept, we summarize the features of the data we use to simulate the housing market in Wake County, North Carolina.

3.1. Characterizing Equilibrium in the Housing Market

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. 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 spatially delineated 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 incomes, 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 home and a quantity of *c* that maximize its utility subject to a budget constraint:

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

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. In order to define this concept more formally, let b_{ij} denote household *i*'s bid for the j^{th} home, and let A_{ij} be an assignment indicator where $A_{ij} = 1$ if and only if household *i* occupies that home. Then 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.[6]

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.

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

This is Rosen's (1974) bid function. It can be used to solve for a locational equilibrium, given a

[6] Equations (2)-(3) are equivalent to the equilibrium concept defined in equations (2)-(4) of CDM.

parametric specification for the utility function, information on preferences and income, and data on housing characteristics. Solutions can be obtained numerically by iterating over the assignment of people to homes until equations (2)-(3) are simultaneously satisfied (Cropper et al., 1988).[7] We follow CDM in using this approach to simulate hedonic equilibria in Wake County, North Carolina.

3.2. Simulating Hedonic Equilibria in Wake County, North Carolina

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 housing sales in this market every year, giving us a large population to draw from. Second, the Wake County Revenue Department develops some of the most comprehensive information on the structural characteristics of individual homes of any metropolitan area in the country.[8] Our data span the years 1992 to 2000, they contain information on more than 104,000 individual housing sales, and they have been used in several previous studies including Fulcher (2003), Pope (2008), and Phaneuf et al. (2008). Third, the data include neighborhood characteristics (median commute time, median household income, and the % under 18 in each census block) and urban amenities that vary within and between neighborhoods (proximity to open space and proximity to shopping centers). Having data on all of these variables allows us to rigorously evaluate the performance of econometric methods which seek to compensate for the researcher's inability to observe all of the neighborhood characteristics and urban amenities

[7] The supplemental appendix provides additional detail on the data, code, and numerical algorithm used to solve for a hedonic equilibrium.
[8] For example, variables include the number of fireplaces and the square footage of garages, decks, basements and attics. To make our simulation more representative of the typical hedonic study, we actually exclude information on basements, decks, and attics, and reduce the resolution of information on garages and fireplaces to dummy variables indicating whether these features are present in each home.

that matter to households.

Table 2 provides summary statistics for the Wake County data. The average home sells for approximately $201,000, it has 2.5 baths, is located on a half acre lot, has a fireplace but lacks a garage, has 1900 square feet of heated living space, is about 10 years old, and is located 4 miles from the nearest park and 8 miles from the nearest shopping center. The neighborhood attributes assigned to each home are defined by the census block in which the home is located. For the average home, median household income in the neighborhood is $68,000, median commute time is 23 minutes, and 27 percent of people living in the neighborhood are under the age of 18. Not surprisingly, many of the spatially delineated variables in our data are highly correlated. For example, distance to the nearest shopping center is highly correlated with commute time ($\rho = .77$) and with distance to the nearest park ($\rho = .73$). Among the structural characteristics, the size of the main heated living area is highly correlated with the presence of a garage ($\rho = .67$) and the number of bathrooms ($\rho = .65$). A full table of correlation coefficients is provided in the supplemental appendix together with maps of the region.

To simulate hedonic equilibria we use our definition for the stock of housing together with a parametric specification for the joint distribution of income and preferences. For each simulation, we represent each household's utility from a vector of housing characteristics, X_j, using one of three parametric specifications: Cobb-Douglas, Translog, or Diewert:

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

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 simulations, only the continuously varying variables are transformed. In other words, for the fireplace and garage dummy variables, we set $g(1)=1$ and $g(0)=0$. Heterogeneous preferences for each housing characteristic are assumed to be independent of income and gamma distributed. Selecting a gamma distribution recognizes that the distribution of preferences may be asymmetric about the mean. This makes it easier to calibrate our simulated equilibria 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 simulation are generated by solving for 100 different hedonic equilibria for each of the three utility functions. On each replication, a sample of households is randomly drawn from the Census income distribution for Wake County under the assumption that people are uniformly distributed within each bin. Then a random sample of homes is drawn from our Wake County database of 104,000 housing transactions. Given a random sample of homes and a random sample of households, we solve for the assignment of households to homes and the vector of equilibrium prices which jointly clear the market for housing.[9] This price vector and the associated housing characteristics comprise the simulated data we use to evaluate alterative specifications for the hedonic price function.

3.3. Key Features of the Experimental Design

Five features of our data generating process offer the potential to inform the current hedonic

[9] An optimization routine was used to calibrate the simulation to approximately reproduce the actual distribution of housing prices in Wake County. Details for this part of the data generating process are provided in the supplemental appendix.

literature. First, in addition to running our simulations for the same Translog and Diewert specifications for utility used by CDM, we repeat the analysis using a Cobb-Douglas utility function. This allows us to test the sensitivity of our results to the presence of interactions between characteristics in the preference function. One might expect simpler utility functions to translate into simpler price functions. This could affect the relative performance of different functional forms.

Second, the scale of our simulation is consistent with recent hedonic applications. Each of our 300 unique hedonic equilibria characterizes a market with 2000 homes, randomly drawn from our Wake population. The resulting simulated data sets are similar in size to the median number of observations (1917) in the 69 published studies that have cited CDM. In comparison, CDM simulated equilibria in a market with 200 homes. Our order of magnitude increase in sample size is important because it permits a practical assessment of the potential bias/efficiency tradeoffs in estimation while simultaneously allowing us to include a realistic set of spatial fixed effects in the regressions.

Third, our simulations depict a spatially explicit urban landscape. Data on five neighborhood attributes and urban amenities are attached to each home in the Wake County data. When we draw a random sample of homes from this database and solve for a hedonic equilibrium, the simulated market will reflect the actual patterns of spatial correlation between neighborhood attributes and housing amenities in Wake County. This realism is particularly important is assessing the consequences of omitting spatially delineated amenities from the price function. We also use the latitude and longitude of each home to develop spatial fixed effects for census tracts and to generate the weights matrices required to estimate spatial lag models and spatial error models (Anselin, 1988).

Fourth, we consider a diversity of omitted variable scenarios. In their original analysis, CDM omitted two housing characteristics—*lot size* and the *number of rooms*. While concern about omitted variables has intensified since their study, the focus has shifted from omitted characteristics of the *home* to omitted characteristics of the *neighborhood*. One reason is that data on structural characteristics have become readily available through "assessor" property value databases.[10] Perhaps more importantly, 85% of the studies in table 1 focus on estimating the MWTP for a spatially delineated amenity such as air quality, school quality, or exposure to hazardous waste. It is often natural to expect these amenities to be correlated with unobserved neighborhood characteristics. Suppose we want to estimate the MWTP for school quality, for example. If homeowners also care about crime rates, and low quality schools tend to be located in high-crime areas, failing to control for crime rates will place an upward bias on the hedonic coefficient for school quality. School quality is one of several amenities that can be homogenous within a neighborhood while varying systematically between neighborhoods. The patterns of spatial correlation between included and omitted neighborhood characteristics may be very different from the spatial correlations between included and omitted structural characteristics. We investigate the consequences of unobserved neighborhood characteristics through two types of omitted variable scenarios.

In the first scenario, we observe all of the structural characteristics of the home but omit the same three spatially delineated attributes from every regression: (1) median income in the census block, (2) share of the population under 18 in the census block, and (3) distance from the home to the nearest shopping center. This "spatial" omitted variable scenario allows us to

[10] 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.

evaluate how the patterns of spatial correlation between observed and unobserved variables ultimately contaminate estimates for MWTP. Moreover, it allows us to measure the relative success of different econometric strategies for purging these sources of contamination. In the second omitted variable scenario, we randomly choose three characteristics to omit from each regression. This “random” approach to selecting variables to omit provides a robustness check on the specific pattern of parameter contamination generated by our “spatial” scenario.

Fifth, we expand the scope of the simulation to evaluate panel data estimators. A recent wave of studies has sought to use temporal variation in the supply of public goods and urban amenities to identify the MWTP. The basic idea is to estimate the rate at which shocks to these amenities are capitalized into housing prices. Implementing this approach requires data from two different equilibria; before and after a shock that “treats” a subset of homes in the market. By collecting data on the same homes at multiple points in time, one can purge the bias that would otherwise arise from time-constant omitted variables. Recent examples of this quasi-experimental approach to hedonic estimation include Davis (2004), Chay and Greenstone (2005), Greenstone and Gallagher (2008), and Pope (2008). We mimic their quasi-experimental designs by simulating the movement from an initial hedonic equilibrium to a new equilibrium following an exogenous shock to the urban landscape. In other words, we generate panel data on equilibrium prices for the same homes before and after a large shock to the market. To make this scenario realistic we base it on an actual shock to commute times in Wake County that occurred following the construction of a new beltline highway in the late 1990s.

Of course it comes as no surprise that panel data estimators are capable of purging the bias from time-constant omitted variables. Rather than simply confirming this general result in the context of the hedonic model, we use our simulations to investigate emerging issues in the quasi-experimental literature. For example, it is not obvious how different panel data estimators will perform in the presence of time-varying omitted variables. Likewise, while large shocks to the market can change the shape of the equilibrium price function over time, hedonic theory does not provide any guidance on how to control for these changes (Epple, 1987). Finally, based on the increased sensitivity of linear panel models to measurement error (compared to linear cross-section models), one might also expect them to be more sensitive to misspecifications in functional form. Our simulations provide the means to investigate these issues in a real urban landscape.

4. Results

In total, we ran 54,000 regressions based on 540 different hedonic models. The models varied according to the shape of the utility function (*Cobb-Douglas*, *Translog*, *Diewert*); the parametric form used to estimate the hedonic price function (*linear*, *log-linear*, *log-log*, *Box-Cox linear*, *quadratic*, *Box-Cox quadratic*); the spatial controls for omitted variables (*no controls*, *fixed effects for census tracts*, *spatial error model*, *spatial lag model*); the approach to panel data estimation (*pooled cross-section*, *differences-in-differences*, *first differences*); the omitted variable scenario (*none*, *spatial*, *random*); and the number of homes in the simulated market (*200*, *2000*). Between 2 and 11 variables entered the price function in each regression, generating more than half a billion estimates for individual MWTP. With this multiplicity of models and wealth of estimates, the clear challenge is to find a coherent way to synthesize the results.

To quantify the performance of an individual model, we first calculate the difference between every household's MWTP for each housing attribute and the corresponding partial derivative of the estimated price function, $\hat{P}(x)$, on each Monte Carlo replication. Equation (9) defines this error, e_{ikr}, in estimating household *i*'s valuation of characteristic *k* on replication *r*.

$$e_{ikr} = \partial \hat{P}_r(x_i) / \partial x_k - MWTP_{jkr} . \tag{9}$$

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

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

The normalized mean and standard deviation are simply the mean ($\bar{e}_{kr}$) and standard deviation ($\bar{s}_{kr}$) of the error from (9), divided by the average MWTP for characteristic *k*. Finally, averaging β_{kr} and S_{kr} over all 100 Monte Carlo replications summarizes the overall success of the model in estimating the MWTP for characteristic *k*.

Even after developing model-specific summary statistics, we still have far too many results to present in a reasonable number of tables. We address this problem in two ways. First, we highlight what we think are the most important results to emerge from the simulation. Part 1 of this section demonstrates that our results are consistent with CDM. Part 2 summarizes our results on using spatial fixed effects to control for omitted variables in cross-section data. Then part 3 summarizes the performance of alternative panel data estimators. Finally, in part 4 we conduct a meta-analysis of models that regresses the values for β_{kr} on dummy variables for modeling choices. This provides a convenient way to detect how specific modeling choices systematically contribute to the empirical hedonic model's ability to identify MWTP.

4.1. Comparison to CDM

Table 3 reports the bias in our estimates for MWTP from each of CDM's six functional forms using their Diewert specification for utility.[11] The numbers in the table represent the mean percentage bias and its standard deviation for each housing characteristic calculated over 100 Monte Carlo replications. First consider the left half of the table, where there are no omitted variables. The bias in estimating MWTP ranges from as low as -0.2% for *garage* in the quadratic model (column 5) to as high as 202% for *acreage* in the log-log model (column 3). All six functional forms do a relatively good job of recovering the MWTP for *bathrooms, garage,* and *commute time,* and a relatively bad job of recovering the MWTP for *fireplace*. For other attributes, the results are less consistent across specifications.

To provide an overall summary of performance, the last three rows of the table report the maximum absolute bias calculated over all 11 attributes, as well as the mean absolute bias, and the mean standard deviation. For example, column (2) reports the results from using a semi-log model. In this case, the estimated implicit hedonic prices for individual characteristics differ from households' true MWTP by 51% on average, with a maximum difference of 92% (for *age*). The average standard deviation on the bias for the semi-log estimator is 1.55.

The last two rows of the table highlight an important bias-variance tradeoff. Moving left to right from column (1) to column (6) increases parametric flexibility in the price function. This decreases the average bias in estimating MWTP, while simultaneously increasing its standard deviation. Both trends are monotonic. The linear model in column (1) has the highest average bias (57%) and the lowest average standard deviation (1.55), whereas the quadratic Box-Cox

[11] The most direct comparison is to tables 2 and 3 in Cropper et al. (1988).

model in column (6) has the lowest average bias (23%) and the highest standard deviation (1.72). Since most hedonic practitioners aim to estimate average MWTP, rather than higher moments of its distribution, we use the maximum bias and the mean bias as our primary criteria for model evaluation.[12] Based on these criteria, the quadratic Box-Cox model outperforms all other functional forms when there are no omitted variables, just as in CDM.

Now consider the right half of table 3. Columns 7 to 12 depict a more realistic scenario where the econometrician is unable to observe some features of neighborhoods that matter to households. We want to estimate MWTP for a particular variable—commute time—but we omit four others from the regression: median household income, % under 18, nearest park, and nearest shopping center. This omission produces a large upward bias in our estimate of the MWTP for commute time for all six functional forms. Yet the bias-variance tradeoff no longer applies. Notice that the three most flexible models produce the largest standard deviation <u>and</u> the largest bias.[13] This is the seminal result from CDM. More flexible functional forms are more sensitive to omitted variables. We now move on to evaluate the performance of econometric strategies for addressing omitted variables directly.

4.2. Do Spatial Fixed Effects Purge the Bias from Omitted Variables?

To evaluate the performance of hedonic models with spatial fixed effects, we concentrate on two spatially delineated variables: *commute time* and *nearest park*. We chose these variables to represent varying degrees of spatial heterogeneity. While actual commute times will vary

[12] 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. A full set of results on the standard deviation are provided in the supplementary appendix.

[13] While this pattern of results is the same as in CDM, our quadratic and quadratic Box-Cox models do not perform quite as badly as theirs. This is partly due to using a larger sample size but may also reflect advances in numerical algorithms used to solve for the Box-Cox parameters. For example, CDM used an early version of Shazam software (ver.6) which had a convergence tolerance (0.01 or 0.001) that is fairly weak compared to current standards.

continuously across homes, our proxy measure from the Census varies discretely across census blocks. Therefore, its results may be relevant for other local public goods and amenities that vary discretely across urban neighborhoods, such as school quality, air quality, and demographic composition (e.g. Black, 1999; Chay and Greenestone, 2005).[14] In contrast, *nearest park* varies continuously within and between census blocks. Thus, its results may apply to other spatially explicit amenities such as distance to the nearest superfund site or distance to the nearest registered sex offender (e.g. Greenstone and Gallagher, 2008; Linden and Rockoff, 2008).

When all variables are observed, MWTP for commute time is slightly overestimated whereas MWTP for the nearest park tends to be underestimated. These results are reported in the first two rows of table 4. Rows 5 and 6 display the corresponding results after omitting the remaining three spatial variables (income, % under 18, and shopping). Spatial correlation between observed and unobserved variables increases estimated MWTP for both *commute time* and *nearest park*. For *commute time* the direction of this effect exemplifies the conventional wisdom about omitted variables in hedonic regression. The omitted variables place an upward bias on its estimated coefficient, confounding our subsequent welfare interpretation.

Interestingly, the direction of the effect for *nearest park* contradicts the conventional wisdom. For 5 out of 6 functional forms, the omitted variable "bias" moves our welfare estimates closer to the true willingness-to-pay! For example, the semi-log model underestimates MWTP for nearest park by 83% when all variables are observed. Omitting the other spatial variables inflates its regression coefficient so that we only underestimate MWTP by 52%. The net effect is a 31% reduction in the absolute welfare bias. This counterintuitive result serves as an important reminder that the true shape of the equilibrium price function is unknown. Failing

[14] Depending on the amenity, the appropriate definition for a neighborhood may be a school district, an air basin, or a residential subdivision.

to specify it correctly can lead to badly biased estimates for MWTP even when omitted variables do not present a confounding influence.

The counterintuitive result for nearest park is interesting, but atypical. Rows 7 and 8 report the maximum bias and the average bias calculated over the 8 variables included in each regression. Both statistics tend to be inflated by omitted variables. In the most extreme case, quadratic Box-Cox, average bias and maximum bias are more than four times as large as in the specification without omitted variables (rows 3 and 4).

The last four rows of table 4 report our results from repeating the omitted variable regressions after adding spatial fixed effects. These "fixed effects" are simply dummy variables for census tracts. A random sample of 2000 homes on the typical Monte Carlo replication contained approximately 100 census tract dummies. These dummies are not perfect controls. They do not control for variation in omitted variables within census blocks. Nor do they control for variation between blocks within a tract. Nevertheless, adding them to the regression produces striking results. The spatial fixed effects nearly eliminate the bias in our estimated MWTP for *commute time*, and they substantially reduce the bias for *nearest park* in the semi-log and linear models.

Rows 4, 8, and 12 contain the main result of this subsection. When three spatially delineated variables are omitted from the regressions, the average welfare bias more than doubles for every one of the six functional forms (comparing rows 4 and 8). Spatial fixed effects for census tracts almost completely purge this bias. That is, the average bias in row 12 is only slightly larger than in row 4. Maximum bias is also substantially reduced. Most important is the dramatic change in performance for the more flexible specifications. Notice that the quadratic Box-Cox and quadratic models now outperform all others in terms of maximum bias (row 11) and average bias (row 12). This finding reverses the seminal result from CDM. *Spatial fixed effects appear to rehabilitate flexible functional forms for hedonic estimation.*

A natural concern is that the pattern of results in table 4 may simply reflect the shape of the utility function or the particular variables we chose to omit from the regressions. Therefore, as a robustness check, we repeat the fixed effects simulation for all three utility functions (Translog, Diewert, and Cobb-Douglas) and 100 different omitted variable scenarios. To avoid inadvertent bias in our experimental design a uniform random number generator was used to select which three variables to omit on each Monte Carlo replication. The omitted variables were drawn from the full set of housing characteristics, neighborhood characteristics, and urban amenities. Table 5 reports the results.

Overall, the two Box-Cox models outperform all others. With spatial fixed effects included in the regression, the quadratic Box-Cox model has the lowest maximum bias for each of the three utility functions. It also has the lowest average bias in the Diewert and Cobb-Douglas versions of the simulation. In the Translog simulation the linear Box-Cox model performs the best. The quadratic model also consistently does well in terms of average bias, while the linear and semi-log models consistently perform the worst. Thus, the ability of spatial fixed effects to "rehabilitate" flexible functional forms for hedonic estimation is robust to the shape of the utility function and to the composition of the set of omitted variables.

4.3. Do Panel Data Estimators Purge the Bias from Omitted Variables?

To generate the data needed to evaluate the performance of panel data estimators, we simulate a shock to the market that could serve as a quasi-experiment. In the pursuit of realism, we base the shock on an actual event that occurred during our study period. The construction of a new beltline highway in the late 1990s dramatically reduced commute times for workers living in selected areas of Wake County. With help from the Wake County planning department, we determined which census tracts were likely to have experienced the largest reductions. Then we divided all of the tracts in the county into four categories, each of which we associate with a constant percentage reduction in commute time: large reductions (50%), moderate reductions (30%), small reductions (15%), and very small reductions (5%). Figure 1 provides a map of these tracts in relation to the new highway. Darker shading indicates greater reductions in commute times for the workers living in the associated tracts.

Next, in order to recognize the difficulty in finding a quasi-experiment that isolates a change in a single variable, we simulate small shocks to three other variables: median household income, % under 18, and nearest park. The resulting correlations between changes in variables over time are much lower than the correlations across space, just as in previous applications. The largest correlation coefficient between the changes in any pair of variables is 0.42, compared to a maximum of 0.77 over the spatial correlations in the initial equilibrium.[15]

Finally, after shocking the urban landscape, we solve for a new hedonic equilibrium. The price of every home changes in the new equilibrium, but not every home changes ownership. We only collect data from the new equilibrium on the homes that changed ownership. This approach mimics the housing market dynamics which generate the data used by the empirical hedonic literature. Put differently, while all property values change over time, it is the subset of movers who negotiate the transactions that ultimately enter the databases we use for econometric analysis. To summarize our panel data generating process: for each of the three utility functions, we have conducted 100 simulations of the movement between hedonic equilibria over time following a large shock to commute time and small shocks to other variables that affect utility.

[15] Complete tables of correlation coefficients are provided in the supplemental appendix.

The pre-shock equilibria are the same equilibria we previously used to evaluate the cross-section models.

We evaluate three approaches to estimation. Our first approach naively pools all of the data. The pooled data are used to estimate the same six price functions tested in the cross-section simulations. These regressions include spatial fixed effects but ignore the temporal dimension of the data. The second estimator is a generalized version of difference-in-differences (DID). This approach involves OLS estimation of the linear, semi-log, and log-log price functions using spatial fixed effects, a time dummy, and interactions between the time dummy and each covariate. These interactions control for the possibility that the marginal implicit prices of the covariates may change as we move between different hedonic equilibria.[16] Our final panel data strategy involves estimating first differenced (FD) versions of the linear, semi-log, and log-log models, using data on repeat sales. Since the FD estimator only uses data on homes that actually sold twice, the sample size is considerably smaller than for the DID and pooled estimators.[17]

Table 6 reports the bias in MWTP from panel data estimation with no omitted variables. These results are averaged over all 300 Monte Carlo replications (100 for each utility specification). Notice that the pooled estimators perform poorly, regardless of functional form. Their absolute bias in measuring MWTP for commute time ranges from 76% to 96%. These figures are 3 to 4 times larger than the corresponding cross-section results in table 3. The large increase in bias reflects the inability of pooled estimation to account for changes in the shape of the price function. More precisely, when commute time decreases, MWTP changes, and so does the shape of the equilibrium hedonic price function. Failing to allow for this adjustment

[16] Epple (1987) illustrates how shocks to a hedonic equilibrium can change the shape of the equilibrium price function and Brookshire et al. (1985) and Costa and Kahn (2003) provide evidence that the implicit price of amenities changes over time.

[17] Approximately 82% of the homes in our simulations are resold, leaving the FD estimator with fewer than half as many observations as the DID and pooled estimators.

introduces a temporal variant of functional form misspecification.

By relaxing the "time-constant price function" restriction, the DID estimators reduce the bias on MWTP for commute time substantially. The resulting range of values for the bias (11% to 39%) is comparable to the cross-section estimates without omitted variables. First-differenced estimation also works well. The FD estimators avoid the need to model changes in the price function by measuring the rate at which the change in commute time is capitalized into sale prices. The resulting range of values for the welfare bias (15% to 34%) is essentially the same as for DID.

To allow for comparison across estimators, the measures for $Max\left|\beta_j\right|$ and $Average\left|\beta_j\right|$ in table 6 are only calculated over the four time-varying variables that enter the FD regressions. Based on these statistics, the DID and FD models clearly dominate pooled estimation. The largest biases for DID and FD across all functional forms are smaller than the smallest biases from pooled estimation. Likewise, conditional on functional form, FD and DID both provide better estimates of the MWTP for commute time and nearest park.

Comparing the FD estimates for MWTP in table 6 with the measures for bias in tables 3, 4, and 5 provides a hedonic example of how panel data estimation can eliminate the bias from time-constant omitted variables. As we would expect, the average bias in the three linear FD regressions is much smaller than the bias from cross-section estimation *with* omitted variables and similar to the bias from cross-section estimation *without* omitted variables.

Next we consider the consequences of omitting time-varying variables. As before, we focus on estimating the MWTP for commute time and nearest park. The two other time-varying variables (median income and % under 18) are omitted. Table 7 compares the three estimators based on the mean and standard deviation of the bias in their MWTP estimates for commute time and nearest park. Not surprisingly, pooled OLS still performs poorly. DID performs better, but the omitted variables clearly hurt its performance. This is especially true for the linear and semi-log models where the bias on MWTP for commute time more than doubles compared to table 6. The bias in the FD estimates also increases substantially. This increase is sufficiently large that the FD estimator which performed quite well in table 6 now performs worse than pooled OLS on nearest park and is dominated by the DID estimator.

Finally, we attempt to control for the effect of time-varying omitted variables in the FD specifications by adding spatial fixed effects. The fixed effects lead to a slight reduction in the average absolute bias on welfare estimates for both amenities. However, notice the large increase in the standard deviation on the bias for commute time. This is because, after time-differencing the data and adding census tract dummies, there is very little within-tract variation left to identify the regression coefficient on commute time. In contrast, the FD estimators do not have the same problem with nearest park because it varies continuously within tracts. This distinction illustrates another interesting bias-variance tradeoff. While spatial fixed effects help to absorb the bias associated with time-constant and time-varying omitted variables, they also absorb some of the variation in the amenity of interest, increasing the standard deviation on welfare measures for quality changes. The DID model avoids the temporal dimension of this problem.

4.4. Seeking Synthesis through Meta-Analysis

The previous sections highlighted key results from a selected subset of our simulations. This limited the scope for comparison. Cross-section estimators were compared with cross-section estimators and panel estimators were compared with panel estimators. Yet one might like to see a comparison *between* cross-section and panel estimators, especially since recent hedonic studies have observed that they tend to have dramatically different implications (e.g. Chay and Greenestone, 2005; Greenstone and Gallagher, 2008; Bayer et al., forthcoming). The need for brevity also prevented us from summarizing results from other interesting specifications. For example, we also tested spatial lag models, spatial error models, nonlinear estimators for panel dataa, and regressions with smaller sample sizes. In addition to omitting potentially useful information, a selective approach to summarizing a specification search can invalidate the statistical properties of the research design (Leamer, 1983). To address these concerns, we adapt the meta-analytic approach to model implementation from Banzhaf and Smith (2007) to synthesize our full set of results, while simultaneously allowing for broader comparison across specifications.

Meta-analysis of model implementation involves regressing model outputs on indicator variables for model inputs. The estimated coefficients summarize the ways in which subjective modeling decisions contribute to the performance of the model. Specifically, the value for the normalized welfare bias (β_{kr}) of each characteristic on each Monte Carlo replication is regressed on indicator variables for: (1) functional form, (2) the type of spatial control for omitted variables, (3) the type of estimator, (4) the size of the market, (5) selected interactions between the first four categories, (6) the omitted variable scenario, (7) the housing characteristic, (8) the shape of the utility function, and (9) the particular sample of homes. Notice that the first five categories describe modeling choices over which the econometrician exerts some control. The last four categories describe features of the simulation; they serve as control variables in the meta-analysis.

Table 8 presents the results. The first column uses data from all 11 housing characteristics, the second column uses data on the 5 spatial characteristics (i.e. census block variables and spatially delineated amenities) and the last column uses data on the commute time variable that was the subject of our simulated quasi-experiment. Our primary interest lies in the second column because 85% of the empirical hedonic studies in our database focus on estimating the MWTP for a single spatially delineated amenity. We include the other two columns because results for the full set of housing characteristics provide the best comparison to CDM and results for commute time are useful in summarizing the performance of alternative estimators when data on the variable of interest are generated by a quasi-experiment. The adjusted R^2 ranges from 0.66 to 0.68 across the three specifications, indicating that hedonic modeling choices explain most of the variation in model performance.[18] These regressions use between 23,537 and 186,946 observations on the bias in measuring MWTP in the presence of omitted variables.

Since all of the regressors are indicator variables, interpretation of results depends on the reference set of modeling choices defined by the excluded indicators. The reference model has a semi-log functional form, is estimated by OLS, uses 200 observations, and has no spatial controls for omitted variables. With this model as the point of reference, the coefficients on the functional form indicators reinforce the main result from CDM. Simpler functional forms perform better in the presence of omitted variables, with the Box-Cox linear and log-log specifications as the top two performers. For example, the Box-Cox linear coefficient of -0.08 in the first column indicates that, all else constant, this model produces estimates for MWTP with 8% less bias than the baseline semi-log model. In comparison, the linear and quadratic models perform the worst.

[18] We dropped 254 observations with biases greater than 500%. These outliers represent approximately one tenth of one percent of our full sample. While dropping these observations does not change the overall pattern of results, it leads to a substantial improvement in model fit. An alternative strategy would be to transform the dependent variable to reduce the influence of outliers. This would allow us to avoid dropping observations but it would complicate interpretation of model coefficients.

Coefficients on the set of indicator variables for market size indicate that the accuracy in estimating MWTP for spatial characteristics is improved by increasing the size of the market. This comes as no surprise. What is surprising is the magnitude of the improvement for the most flexible functional forms. Increasing the sample size from 200 to 2000 produces a 17% reduction in the bias for the quadratic model (4.5% + 12.5%) and an 11% reduction in the bias for the Box-Cox quadratic model (4.5% + 6.8%). Equally important, increasing the sample size enables the use of spatial fixed effects for census tracts without concern for the loss in degrees of freedom.

Spatial fixed effects are clearly the preferable strategy for addressing spatially correlated omitted variables in cross-section data. Adding fixed effects to the model reduces the bias by 7.3% to 8.6% depending on whether we focus on all characteristics or just the subset of spatial characteristics. Interestingly, the spatial error model provides only a marginal improvement over a model with no spatial controls, while the spatial lag model actually performs worse. This relatively poor performance probably reflects the complexity of spatial correlations in our data. With the large number of census tracts in our market, spatial fixed effects offer a fairly flexible means of absorbing the irregular concentration of omitted variables. In contrast, the spatial lag and spatial error models impose a rigid structure on the spatial relationships between included and omitted variables. It is natural to expect this rigidity to add to the bias in measuring MWTP. Nevertheless, it is important to point out that our assessment of the two spatial econometric models relies on the "nearest neighbor" specification for the weights matrix. No formal theory exists for the appropriate construction of this matrix and it is possible that a different specification may lead to materially different results. This is a potentially interesting topic for future research.

The final set of indicator variables in table 8 describes the relative performance of different cross-section and panel data estimators. Compared to cross-section estimation, naively pooling panel data increases the average bias in measuring MWTP for spatial characteristics by 4.6%, despite the increase in sample size. In comparison, the difference-in-differences estimator produces a substantial improvement (12.3%). The stark difference between pooled estimation and DID highlights the importance of controlling for changes in the shape of the equilibrium price function over time. The consequences of failing to do so are the most severe for the variable that changed the most—commute time. In this case, pooled OLS estimates for MWTP are 24% more biased than the DID estimates. The empirical importance of this result is reinforced by the recent wave of quasi-experimental studies which have demonstrated that the public goods we seek to assign implicit prices to often change substantially over time (e.g. Chay and Greenstone, 2005; Greenstone and Gallagher, 2008; Linden and Rockoff, 2008; Pope, 2008).

The first-differenced model avoids the need to model changes in the shape of the equilibrium price function by estimating the rate at which changes in housing characteristics are capitalized into housing prices. The resulting estimates of the MWTP for commute time in the log-log version of the model are measured with comparable accuracy to the DID estimates (12.9% and 13.3% less bias than OLS estimation of the semi-log model). While the linear and semi-log FD models do not perform quite as well, they still do much better than pooled OLS.

As a final panel data test, we attempted linear Box-Cox estimation of the first-differenced model. This was an utter failure. Our meta-analytic regression predicts a welfare bias on commute time that is 35% higher than the baseline OLS model. It seems likely that differencing the data simply removes the spatial variation needed to pin down the parameters of the nonlinear model. This finding seems analogous to the main result from CDM's original study: in the presence of *time-varying* omitted variables, the more parsimonious *panel data* estimators *with spatial fixed effects* tend to perform the best.

5. Conclusions

Omitted variables have long been viewed as a hidden danger in hedonic modeling and this perception has driven the econometric evolution of the literature. Since Cropper, Deck, and McConnell's (1988) influential study, researchers have been willing to make untested assumptions about the shape of the hedonic price function in order to hedge against the risk of omitted variable bias. While this tradition continues, the literature has also evolved to address the problem directly. It has become routine to use spatial fixed effects to help absorb the confounding influence of omitted variables and recent studies have used quasi-experimental methods to purge time-constant omitted variables from the model. Our simulation results provide the first hard evidence that these new methods work. The evidence also suggests it is time to rethink some of the stylized facts about hedonic estimation.

One stylized fact is that more parsimonious functional forms are less susceptible to omitted variable bias. This is the seminal result from Cropper, Deck, and McConnell (1988). We have replicated their result and we find that it no longer holds when spatial fixed effects are added to the model. When spatial fixed effects are used to control for omitted variables, we find that "flexible" specifications for the price function, such as the quadratic Box-Cox model, outperform the simpler linear, log-linear, and log-log specifications that have dominated empirical practice for the past two decades. This result is robust to features of our simulation including the specification for utility that underlies hedonic equilibrium and the composition of the set of omitted variables. Thus, it may be time to reconsider the quadratic Box-Cox model and other flexible specifications for empirical hedonic research.

Another stylized fact is that amenities have time-constant implicit prices. Just over half of the 85 studies we reviewed use more than five years of sales data. They control for changes in amenities over time but, with few exceptions, they restrict the implicit prices of those amenities to be time-constant. This restriction is not supported by hedonic theory. In general, shocks to preferences, income, information, or to the amenities themselves can change the shape of the equilibrium price function (Epple, 1987; Rosen, 2002). Ignoring this adjustment can bias welfare measures. We find this to be important. Simply pooling data before and after shocks to neighborhood amenities conveys a bias in our welfare estimates that is similar in magnitude to the bias from omitted variables. Fortunately, we also find that this bias can be averted by using a generalized difference-in-differences estimator that includes interactions between a time dummy and all of the characteristics of the hedonic price function. A first-differences approach is also effective. These results are encouraging, but it is important to acknowledge that they describe a fairly "clean" quasi-experiment where a few variables receive small to moderate shocks and markets instantly clear. Real housing markets may experience irregular shocks of varying magnitudes with slow transition paths. Therefore, an important topic for future research will be to extend hedonic theory and econometrics to investigate what the complexity of real world housing market dynamics can reveal about the willingness-to-pay for changes in nonmarket goods and services.

There are many other interesting results to emerge from our simulations. For example, we find that spatial error models and spatial lag models have particular difficulty in dealing with time-constant omitted variables. We also find that difference-in-differences and first-differenced specifications for the price function can perform quite well in the presence of *time-varying* omitted variables as long as we include spatial fixed effects in the regression. The challenge lies in determining their geographic resolution. If the fixed effects are too large, they may fail to absorb meaningful variation in the omitted variables. If they are too small, they may absorb most of the variation in the amenity of interest. Evaluating this bias/efficiency tradeoff would be another interesting direction for further research.

In retrospect, we were motivated to undertake this research by the rapid evolution of econometrics in the hedonic literature between 1988 and 2008 and by the lack of evidence on the performance of the new methods. If the past is any indication of the future, hedonic econometrics will continue to evolve in new and exciting ways in the upcoming years. People are just beginning to investigate the connections between structural and reduced form estimation (Bayer et al. 2007), regression discontinuity designs (Greenstone and Gallagher, 2008), and the importance of market friction due to moving costs (Bayer et al. forthcoming) for example. Our hope is that the future will bring a tighter feedback loop between the development and application of new estimators and simulation-based evaluation of their performance. We have taken a first step in this direction.

References

Anselin, Luc. 1988. *Spatial Econometrics: methods and models*, Dordecth: Kluwer Academic.

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

Banzhaf, Spencer H. and V. Kerry Smith. 2007. "Meta Analysis in Model Implementation: Choice Sets and the Valuation of Air Quality Improvements." *Journal of Applied Econometrics*, 22(6): 1013-31.

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, Nathaniel Keohane, and Christopher Timmins. in press. "Migration and Hedonic Valuation: The Case of Air Quality." *Journal of Environmental Economics and Management*.

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

Brookshire, David S., Mark A. Thayer, John Tschirhart, and William D. Schulze. 1985. "A Test of the Expected Utility Model: Evidence from Earthquake Risks." *Journal of Political Economy*, 93(2): 369-89.

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.

Cropper, Maureen L., Leland Deck, Nalin Kishor, and Kenenth E. McConnell. 1993. "Valuing Product Attributes Using Single-Market Data: A Comparison of Hedonic and Discrete-Choice Approaches." *Review of Economics and Statistics*, 75(2): 225-32.

Costa, Dora L and Matthew E. Kahn. 2003. "The Rising Price of Nonmarket Goods." *American Economic Review*, 93(2): 227-32.

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.

Fulcher, Charles M. 2003. "Spatial Aggregation and Prediction in the Hedonic Model." Ph.D. Dissertation, North Carolina State University (Economics).

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.

Greenstone, Michael and Justin Gallagher. 2008. "Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program." *Quarterly Journal of Economics*, 123(3): 951-1003.

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

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review*, 73(1): 31-43.

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.

Linden, Leigh and Jonah E. Rockoff. 2008. "Estimates of the Impact of Crime Risk on Property Values from Megan's Laws." *American Economic Review*, 98(3): 1103-27.

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, Raymond B. 2005. "Property Value Models," in *Handbook of Environmental Economics, Volume 2*. Karl-Göran Mäler and Jeffery Vincent eds. Amsterdam: North Holland Press.

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. 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.

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

Rosen, Sherwin. 2002. "Markets and Diversity." *American Economic Review*, 92(1): 1-15.

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 Studies: 1998-2008 [a]

		CDM	SSCI (69)	ECON (16)
Functional form	main specification: lin-lin, log-lin, log-log		80%	94%
	main specification: Box-Cox		17%	6%
Sample Size	Median # observations	200	1,917	10,599
	published in 1989-1998		593	934
	published in 1999-2008		2,459	13,414
	Distribution of studies by # observations			
	0 to 200	•	6%	0%
	201 to 500		10%	13%
	501 to 1,000		23%	19%
	1,001 to 10,000		39%	19%
	more than 10,000		22%	50%
Market Dimenions	Geography			
	smaller than a city		7%	0%
	city or county	•	42%	25%
	multiple cities or counties		42%	38%
	nation		9%	38%
	Time Period			
	0 to 1 year		7%	0%
	1 to 2 years	•	22%	13%
	2 to 5 years		26%	13%
	5 to 10 years		28%	31%
	more than 10 years		17%	44%
Space and Time	spatial error or spatial lag model	no	14%	0%
	quasi-experimental identification	no	20%	69%
	spatial variable of interest	no	84%	88%
	includes temporal fixed effects	no	39%	50%
	day, month, or quarter		14%	25%
	year		25%	25%
	includes spatial fixed effects	no	58%	75%
	neighborhood		23%	38%
	city or county		29%	13%
	region		6%	25%

[a] SSCI (69) is the set of 69 property value hedonic studies that cite CDM according to the Social Science Citation Index and were published between November 1988 and November 2008. It contains 35 journals with an average of 2 papers per journal. ECON (16) is the set of 16 property value hedonic studies published in *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and the *Review of Economics and Statistics*. These studies may or may not cite CDM. They have an average of 4 papers per journal.

TABLE 2—Summary Statistics for Wake County Data [a]

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

[a] *Median household income*, *median commute time*, and *% under 18* are defined for the Census block in which the home is located. *Nearest park* and *nearest shopping center* are uniquely defined for each home.

TABLE 3—Comparison of Basic Results for Alternative Functional Forms to Cropper et al. (1988)

Mean Error / Mean True Price

(Standard Deviation of Error / Mean True Price)

	(1) Linear	(2) Semi-Log	(3) Log-Log	(4) Box-Cox Linear	(5) Quadratic	(6) Box-Cox Quadratic	(7) Linear	(8) Semi-Log	(9) Log-Log	(10) Box-Cox Linear	(11) Quadratic	(12) Box-Cox Quadratic
bathrooms	0.051	-0.120	-0.185	-0.190	-0.408	-0.236	0.825	0.573	0.674	0.651	0.500	0.607
	(1.406)	(1.412)	(1.404)	(1.417)	(1.553)	(1.527)	(1.406)	(1.472)	(1.459)	(1.558)	(1.666)	(1.777)
acreage	-0.204	-0.313	2.020	1.788	0.260	-0.273	-0.225	-0.354	1.183	0.861	0.101	-1.070
	(1.530)	(1.528)	(2.180)	(2.122)	(1.517)	(1.817)	(1.530)	(1.527)	(1.823)	(1.715)	(1.543)	(2.730)
garage	-0.201	0.352	0.072	0.334	-0.002	0.068	0.539	1.050	0.809	1.194	0.605	0.930
	(1.450)	(1.464)	(1.446)	(1.544)	(1.520)	(1.530)	(1.450)	(1.540)	(1.509)	(1.824)	(1.528)	(1.756)
fireplace	-1.197	-0.557	-0.937	-0.577	-0.537	-0.367	-0.696	-0.081	-0.412	0.096	-0.203	0.278
	(1.432)	(1.418)	(1.430)	(1.424)	(1.541)	(1.474)	(1.432)	(1.424)	(1.418)	(1.500)	(1.685)	(1.961)
main heated living area	0.639	0.330	0.107	-0.016	-0.064	-0.147	2.184	1.698	1.689	1.550	1.734	1.766
	(1.407)	(1.455)	(1.412)	(1.423)	(1.526)	(1.567)	(1.407)	(1.670)	(1.551)	(1.735)	(1.615)	(1.990)
age	-0.948	-0.924	-0.130	-0.204	-0.757	-0.068	-0.921	-0.902	-0.314	-0.474	-0.806	-0.206
	(1.965)	(1.960)	(1.841)	(1.865)	(1.939)	(1.865)	(1.965)	(1.958)	(1.825)	(1.853)	(1.944)	(1.897)
median household income	-0.318	-0.406	-0.488	-0.477	-0.476	-0.443	This variable has been omitted					
	(1.221)	(1.214)	(1.201)	(1.200)	(1.287)	(1.293)						
median commute time	0.260	0.269	0.271	0.374	0.306	0.252	1.003	1.123	1.397	1.761	1.430	1.622
	(1.808)	(1.811)	(1.816)	(1.900)	(2.016)	(2.048)	(1.808)	(1.884)	(1.988)	(2.372)	(1.875)	(2.399)
% under 18	-0.790	-0.803	-0.651	-0.676	-0.255	-0.384	This variable has been omitted					
	(1.312)	(1.303)	(1.288)	(1.291)	(1.415)	(1.351)						
nearest park	-0.939	-0.834	-0.337	-0.352	-0.477	0.046	This variable has been omitted					
	(1.718)	(1.709)	(1.658)	(1.675)	(1.709)	(2.002)						
nearest shopping center	-0.700	-0.681	0.045	0.064	-0.097	0.263	This variable has been omitted					
	(1.810)	(1.795)	(1.837)	(1.858)	(1.778)	(2.398)						
Maximum $\lvert \beta_k \rvert$	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 $\lvert \beta_k \rvert$	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 4—Cross-Section Hedonic Results with and without Spatial Fixed Effects [a]

	Metric for Comparison	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
				functional form			
			No omitted variables, no spatial fixed effects				
(1)	$\beta_{\text{commute time}}$	0.26	0.27	0.27	0.37	0.31	0.25
(2)	$\beta_{\text{nearest park}}$	-0.94	-0.83	-0.34	-0.35	-0.48	0.05
(3)	Max $\lvert \beta_j \rvert$	1.20	0.92	2.02	1.79	0.76	0.37
(4)	Average $\lvert \beta_j \rvert$	0.40	0.34	0.37	0.35	0.26	0.13
			3 omitted variables, no spatial fixed effects*				
(5)	$\beta_{\text{commute time}}$	0.61	0.64	0.90	1.16	0.96	1.18
(6)	$\beta_{\text{nearest park}}$	-0.61	-0.52	0.11	0.12	-0.20	0.40
(7)	Max $\lvert \beta_j \rvert$	2.14	1.65	1.81	1.39	1.52	1.54
(8)	Average $\lvert \beta_j \rvert$	0.81	0.71	0.78	0.76	0.64	0.71
			3 omitted variables, spatial fixed effects*				
(9)	$\beta_{\text{commute time}}$	-0.08	0.08	0.06	0.22	0.09	0.21
(10)	$\beta_{\text{nearest park}}$	-0.41	-0.29	0.27	0.34	-0.19	0.35
(11)	Max $\lvert \beta_j \rvert$	0.88	0.86	1.68	1.32	0.75	0.65
(12)	Average $\lvert \beta_j \rvert$	0.43	0.38	0.40	0.41	0.28	0.26

[a] All results are based on the Diwert specification for utility and averaged over 100 Monte Carlo replications. Omitted variables (1) median household income, (2) % under 18, (3) distance to the nearest shopping center.

TABLE 5—Cross-Section Hedonic Results with Spatial Fixed Effects: Robustness Checks [a]

Metric for Comparison	Linear	Semi-Log	Log-Log	Box-Cox Linear	Quadratic	Box-Cox Quadratic
			functional form			
			Diewert Utility Function			
Max | β_j |	0.88	0.86	1.52	0.90	0.75	0.64
Average | β_j |	0.50	0.46	0.42	0.38	0.35	0.31
			Translog Utility Function			
Max | β_j |	0.94	0.93	0.70	0.69	0.86	0.69
Average | β_j |	0.56	0.52	0.25	0.23	0.42	0.31
			Cobb-Douglas Utility Function			
Max | β_j |	1.15	0.96	1.08	0.88	0.91	0.74
Average | β_j |	0.66	0.61	0.47	0.42	0.49	0.35

[a] Every model includes spatial fixed effects for census tracts. All results are averaged over 100 Monte Carlo replications. Omitted variables: three were randomly selected on each replication.

TABLE 6—Panel Estimation Hedonic Results with No Time-Varying Omitted Variables [a]

	Pooled OLS						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
$\lvert \beta_{\text{commute time}} \rvert$	0.82	0.76	0.87	0.89	0.91	0.96	0.19	0.11	0.39	0.33	0.34	0.15
$\lvert \beta_{\text{nearest park}} \rvert$	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 $\lvert \beta_j \rvert$	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 $\lvert \beta_j \rvert$	0.64	0.64	0.60	0.61	0.66	0.59	0.40	0.40	0.38	0.35	0.38	0.27
fixed effects	x	x	x	x	x	x	x	x	x			
average N	3643	3643	3643	3643	3643	3643	3643	3643	3643	1643	1643	1643

[a] Results for each model are averaged over 300 Monte Carlo replications, 100 for each of the Cobb-Douglas, Diewert, and Translog specifications for utility. Omitted variables: none. Time-varying variables: (1) commute time, (2) nearest park, (3) % under 18, and (4) distance to nearest shopping.

TABLE 7— Panel Estimation Hedonic Results with Time-Varying Omitted Variables [a]

	Pooled OLS			Difference-in-Difference			First Difference					
	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log	Linear	Semi-Log	Log-Log
$\lvert \beta_{\text{commute time}} \rvert$	0.69	0.66	0.81	0.41	0.33	0.43	0.56	0.52	0.49	0.44	0.58	0.38
	(0.07)	(0.07)	(0.05)	(0.09)	(0.09)	(0.07)	(0.05)	(0.05)	(0.05)	(0.35)	(0.29)	(1.35)
$\lvert \beta_{\text{nearest park}} \rvert$	0.62	0.60	0.18	0.54	0.52	0.15	0.84	0.81	0.31	0.73	0.76	0.18
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.06)	(0.03)	(0.03)	(0.06)	(0.05)	(0.04)	(0.07)
fixed effects	x	x	x	x	x	x				x	x	x
average N	3643	3643	3643	3643	3643	3643	1643	1643	1643	1643	1643	1643

[a] Results for each model are averaged over 300 Monte Carlo replications, 100 for each of the Cobb-Douglas, Diewert, and Translog specifications for utility. Omitted variables (1) median household income, (2) % under 18, (3) distance to the nearest shopping center. Time-varying variables: (1) commute time, (2) nearest park, (3) % under 18, and (4) distance to the nearest shopping.

TABLE 8— Meta Analysis of Hedonic Results

Modeling Choice	All Characteristics	All Spatial Characteristics	Median Commute Time
Functional form			
linear	0.0614***	0.0220***	0.0180
log-log	-0.0711***	-0.1325***	0.1819***
box-cox linear	-0.0829***	-0.1256***	0.2531***
quadratic	0.0291***	-0.0174*	0.2947***
box-cox-quadratic	-0.0496***	-0.0937***	0.3100***
Market Size			
market size = 2000	0.0026	-0.0454***	-0.0507***
quadratic, 2000	-0.1253***	-0.1248***	-0.1700***
box-cox-quadratic, 2000	-0.1199***	-0.0685***	-0.0986***
Spatial controls			
spatial fixed effects	-0.0861***	-0.0730***	-0.1482***
spatial error model	0.1647***	-0.0181*	0.0183
spatial lag model	0.1833***	0.0300***	0.1120***
Estimator			
pooled OLS / ML	0.0030	0.0465***	0.1119***
difference-in-differences	-0.0485***	-0.1229***	-0.1332***
first difference (FD)	0.1599***	0.1002***	0.0498***
linear, FD	-0.0877***	-0.0542***	-0.0550*
log-log, FD	-0.1798***	-0.1242***	-0.1789***
box-cox linear, FD	0.2468***	0.3042***	0.3496***
N	186,946	68,152	23,537
R^2 adjusted	0.6568	0.6778	0.6708

legend: * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Figure 1: Location of Interstate Project Relative to Wake County Census Tracts