LOCATION CHOICE AND THE VALUE OF SPATIALLY DELINEATED AMENITIES

by

Kelly C. Bishop

Department of Economics
Duke University

Date: ______________________________
Approved:

____
Christopher Timmins, Supervisor

____
Peter Arcidiacono

____
Patrick Bayer

____
Martin Smith

____

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Economics
in the Graduate School of
Duke University

2008

ABSTRACT

LOCATION CHOICE AND THE VALUE OF SPATIALLY DELINEATED AMENITIES

by

Kelly C. Bishop

Department of Economics
Duke University

Date: ______________________

Approved:

____
Christopher Timmins, Supervisor

____
Peter Arcidiacono

____
Patrick Bayer

____
Martin Smith

____

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics in the Graduate School of Duke University

2008

Abstract

In the first chapter of this dissertation, I outline a hedonic equilibrium model that explicitly controls for moving costs and forward-looking behavior. Hedonic equilibrium models allow researchers to recover willingness to pay for spatially delineated amenities by using the notion that individuals "vote with their feet." However, the hedonic literature and, more recently, the estimable Tiebout sorting model literature, has largely ignored both the costs associated with migration (financial and psychological), as well as the forward-looking behavior that individuals exercise in making location decisions. Each of these omissions could lead to biased estimates of willingness to pay. Building upon dynamic migration models from the labor literature, I estimate a fully dynamic model of individual migration at the national level. By employing a two-step estimation routine, I avoid the computational burden associated with the full recursive solution and can then include a richly-specified, realistic state space. With this model, I am able to perform non-market valuation exercises and learn about the spatial determinants of labor market outcomes in a dynamic setting. Including dynamics has a significant positive impact on the estimates of willingness to pay for air quality. In addition, I find that location-specific amenity values can explain important trends in observed migration patterns in the United States.

The second chapter of this dissertation describes a model which estimates willingness to pay for air quality using property value hedonics techniques. Since Rosen's seminal 1974 paper, property value hedonics has become commonplace in the non-market valuation of environmental amenities, despite a number of well-known methodological problems. In particular, recovery of the marginal willingness to pay function suffers from important endogeneity biases that are difficult to correct with instrumental variables procedures [Epple (1987)]. Bajari and Benkard (2005) propose a

"preference inversion" procedure for recovering heterogeneous measures of marginal willingness to pay that avoids these problems. However, using cross-sectional data, their approach imposes unrealistic constraints on the elasticity of marginal willingness to pay. Following Bajari and Benkard's suggestion, I show how data describing repeat purchase decisions by individual home buyers can be used to relax these constraints. Using data on ozone pollution in the Bay Area of California, I find that endogeneity bias and flexibility in the shape of the marginal willingness to pay function are both important.

Finally, in the third chapter of this dissertation, I combine the insights of the Bajari-Benkard inversion approach employed in second chapter with more standard estimation techniques (*i.e.*, Rosen (1974)) to arrive at a new hedonic methodology that allows for flexible and heterogeneous preferences while avoiding the endogeneity problems that plague the traditional Rosen two-stage model. Implementing this estimator using the Bay Area ozone data, I again find evidence of considerable heterogeneity and of endogeneity bias. In particular, I find that a one unit deterioration in air quality (measured in days in which ozone levels exceed the state standards) raises marginal willingness to pay by \$145.18 per year. The canonical two-stage Rosen model finds, counter-intuitively, that this same change would reduce marginal willingness to pay by \$94.24.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my advisor, Chris Timmins, for his advice and encouragement during the writing of this dissertation. He has taught me a great deal about economic research, the profession, and life in general. The second and third chapters of this dissertation are based on collaborative work with Chris.

I am grateful to my committee members, Peter Arcidiacono, Pat Bayer, and Marty Smith, for their friendship and professional guidance throughout this entire process, the members of the Applied Micro group at Duke for their helpful comments, Jon James for his GIS assistance, and Nat Keohane for sharing his PM10 data.

I would also like to thank my mother, Mary Bishop, my grandparents, Catherine and Paul Donelan, and all of my aunts, uncles, and cousins for their love and encouragement during the many years of my education. There is no better group of cheerleaders.

Most importantly, I would like to thank my husband, Alvin Murphy. As both a loving spouse and an extremely capable economist, he has been an unfailing source of support and guidance.

Chapter 1

A Dynamic Model of Location Choice and Hedonic Valuation

1.1 Introduction

Since the seminal theoretical models of Hicks (1932) and Sjaastad (1962), economists have been interested in examining both the determinants and consequences of individuals' location decisions. Individuals choose their location for a variety of reasons, including employment opportunities and family-related motivations. Importantly, they also care about local public goods and amenities when making these decisions. This feature of their behavior provides a basis for the non-market valuation of local attributes using the ideas in Tiebout (1956).[1] Complicating such an exercise, however, is the fact that location choice is also an inherently dynamic decision process. Individuals face high costs associated with moving. We therefore expect that they would look to the future with regard to wage opportunities and time-varying location attributes when choosing where to live today. The hedonic and empirical Tiebout sorting literatures have, however, essentially ignored these complications. In the case of amenity valuation, ignoring dynamic considerations would bias estimates of willingness to pay downwards if individuals chose locations with high predicted, but low current levels of the amenity in question. Similarly, ignoring moving costs would also bias estimates downwards if individuals had ties, either financial or psychological, to areas with low amenity levels. My analysis addresses these shortcomings by developing and estimating a fully dynamic model of national migration and using it to recover individual

[1] Notable papers which incorporate these ideas into hedonic and sorting models include: Rosen (1979), Roback (1982), Hoehn, Berger, and Blomquist (1987), Blomquist, Berger, and Hoehn (1988), Epple and Sieg (1999), Sieg, Smith, Banzhaf, and Walsh (2004), Bajari and Kahn (2005), Bayer, McMillan, and Rueben (2005), and Kuminoff (2007).

willingness to pay for air quality. Ignoring dynamic concerns is shown to significantly downwardly bias these estimates.

Because of the computational burden associated with the traditional, full-solution method of dynamic analysis, previous dynamic models of migration have been forced to either limit the individual to a simple decision of move-stay, which is inappropriate for valuing amenities that vary spatially, or to essentially ignore the role of local amenity values in migration altogether. In a model of marital status and family location decisions, Gemici (2007) specifies the geographic choice set as the nine Census divisions in the United States. At this level of geography, it is impractical to infer preferences for spatially varying amenities. Kennan and Walker (2006) estimate a model of expected income and individual location decisions, allowing the choice set to be defined at the finer level of U.S. states. However, the computational burden of the traditional estimator limits the number and the type of parameters that can be estimated, particularly as the number of elements in the choice set grows. While including mean wages (deflated to reflect differences in cost of living), population, and historical averages of temperature to describe location amenities, Kennan and Walker specify these attributes as being fixed through time and do not control for unobserved location attributes.

I model the location decisions of individuals using panel data from the National Longitudinal Survey of Youth (NLSY79), explicitly controlling for costs of moving, which are allowed to vary with a variety of factors, including age. Individuals are forward-looking and choose the location that maximizes the discounted stream of benefits in expectation. Wages are allowed to evolve stochastically and the notion of job search is included,[2] as individuals learn their full wage in a given location only after moving there and paying the associated moving costs.

[2]Kennan and Walker allow the location-specific match component of earnings to evolve stochastically. I additionally allow the location-specific mean wage to evolve stochastically, as well as the location-specific cost of living, proxied by median house prices.

To properly define local attributes, one must move to an even finer level of geography than the U.S. state. I estimate my model at the level of metropolitan area,[3] allowing individuals to be forward-looking with respect to a rich set of both fixed and time-varying individual and location attributes. I circumvent the computational burden of the traditional estimation routine by employing the two-step estimator proposed by Arcidiacono and Miller (2007). This method avoids the need for the full recursive solution by estimating the components of the value function in two stages. Based on the conditional choice probability estimation of Hotz and Miller (1993), Arcidiacono and Miller show how to write lifetime utility as a function of current flow utility and the discounted conditional probabilities of choosing particular locations in the future. This simplifies the main estimating equation to one that is linear in the structural parameters and that can handle a large descriptive state space.

The use of Hotz and Miller-style two-step approaches has become increasing popular, particularly to estimate dynamic games in Industrial Organization, as they dramatically reduce the computational burden of these models.[4] The two-step approach in a migration model uses the fact that individuals optimally choose where to live and work based on both current and expected future local attributes. Thus, the probability of choosing a particular location at a given time incorporates the full information set available that period. Therefore, it is possible to replace the difficult terms within the future value component of lifetime utility with future conditional choice probabilities. In the first step of the estimation routine, I flexibly estimate the reduced-form future probability of choosing a particular location, conditional on the agent being at any one of the potential points in

[3]To avoid issues with sparse data, I limit my analysis to the fifty most populous metropolitan statistical areas (MSAs) in the NLSY79. I additionally create nine Census division, non-MSA "catch-alls" to limit attrition from the panel. Thus, the choice set is comprised of fifty-nine locations. With more data describing migration behavior (see footnote 18), my proposed model could handle a much larger choice set.

[4]See, for example, Aguirregabiria and Mira (2007), Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky and Berry (2005), and Pesendorfer and Schmidt-Dengler (2007).

the state space. The remaining structural parameters are then estimated in the second stage, after plugging in the conditional choice probabilities and integrating out over the transitions of the stochastic variables.

This paper makes several contributions to the literature. While previous dynamic models of location choice were forced to essentially ignore the role of amenity values, the use of a two-step approach allows me to include a rich descriptive set of local attributes that are allowed to transition through time, in addition to location fixed-effects (which control non-parametrically for all time-invariant local attributes). This extends the existing labor literature of migration models and allows me to estimate willingness to pay for non-market amenities which vary spatially. Thus, the model extends the existing literature on Tiebout sorting models, which typically ignore the costs associated with migration (both financial and psychological)[5] as well as the forward-looking behavior that individuals exercise in making location decisions. I find that the inclusion of moving costs and forward-looking behavior has a significant impact on estimates of willingness to pay. I estimate a willingness to pay for a one unit decrease in air pollution (measured as a one microgram per cubic meter of air decrease in particulate matter) of \$198.22,[6] which implies an elasticity for the average individual of -0.17.[7] This willingness to pay is 2.2 times larger than that estimated by the dynamic model without ties to birth location and 2.6 times larger than that estimated by a simple static model.[8]

[5] An exception is Bayer, Keohane, and Timmins (2006), which allows for moving costs that depend only on an individual's birth location, but which ignores other types of dynamics. Deriving a willingness to pay for air quality using a static Tiebout sorting model, they find that ignoring moving costs substantially under-estimates the value of clean air.

[6] The willingness to pay is reported in inflation-adjusted 2000 dollars. For ease of comparison to previous estimates (see, for example, Chay and Greenstone (2005) and Bayer, Keohane, and Timmins (2006)), which are expressed in constant 1982-1984 dollars, divide all figures by 1.72.

[7] The elasticity is calculated at a mean income of $\$32,997.26$ (in 2000 dollars) and a mean particulate matter concentration of 27.79 $\mu g/m^3$.

[8] See Smith and Huang (1995) for an excellent review of the willingness to pay for air quality literature.

This paper is organized as follows. I describe my model in Section 1.2. Sections 1.3 and 1.4 discuss the main data sources and the two-step estimation routine in detail. I present the results in Section 1.5. Extensions are briefly described in Section 1.6. Finally, Section 1.7 concludes.

1.2 Model

I model the location decisions of individuals in a finite-horizon framework. In each period, individuals receive flow utility associated with their current location and incur a moving cost if they decide to relocate in the following period. Locations will be defined as a set of U.S. metropolitan areas.

The timing of the decision process is important and highlights the effect of expectations on the location decision. The decision period will be two years in length.[9] An agent begins period t in a location denoted j. The agent knows the full utility flow associated with their current location. If they choose to move (*i.e.* begin period $t+1$ in a new location k), they pay all associated moving costs (which are known in full) in the current period, t. However, the individual only has expectations over the value of living in k, which will not be realized until period $t+1$, after all costs have been paid.

Agents receive flow utility from income, living in the location or region of their birth, the amenities associated with their current location, and an additively-separable choice-specific shock. Moving costs are specified as a function of the agent's age, the distance between the locations, and whether or not the agent has previously lived in the chosen location. Lifetime utility is given by current flow utility and the discounted stream of expected future per-period utilities. Individuals are forward looking with respect to income and time-varying location attributes. Uncertainty comes in the form of the transition of these variables, a location- and individual-specific match component of income, and an idiosyncratic shock

[9] In my primary data source, the NLSY79, individuals are interviewed biennially.

to utility.

1.2.1 Income

In each location in each period, an individual receives the location- and time-specific mean income, a return to their individual characteristics (including age), and an idiosyncratic error on income. Following Kennan and Walker (2006), I assume that this error term can be divided into three distinct components: a fixed location-specific match component, an individual fixed effect, and a transitory component. Thus, the income of individual i, in location j, at time t is specified as:[10]

$$inc_{i,j,t} \ = \ \omega_i'\gamma \ + \ f(age_{i,t}) \ + \ \mu_{j,t} \ + \ (\theta_{i,j} + \eta_i + e_t)$$

where ω_i is the vector of fixed characteristics of individual i, $age_{i,t}$ is individual i's age in period t, and $\mu_{j,t}$ is the location-specific mean wage at time t. The error term is comprised of the match component $\theta_{i,j}$, the individual fixed effect η_i, and the transitory earnings component e_t. All three are assumed to be i.i.d. (across individuals, locations, and time) normally distributed with mean 0 and respective variances σ_θ^2, σ_η^2, and σ_e^2.

Individuals know their individual fixed effect and the current values of their location-specific match component and time-specific transient component. However, an individual does not know the value of their match component in other locations or the value of the transient component in future periods (but does know the distributions from which they are drawn). I additionally allow individuals to know the value of their match component in their prior location.[11]

Income plays an important role in the decision to migrate. However, only the compo-

[10] In this specification, the return to individual attributes is fixed across locations. In practice, I run the model on a homogenous sample, so these effects are captured non-parametrically by the location-specific mean wage. Individual characteristics (besides age, but including education) are held fixed through time.

[11] This introduces "memory" in the form of Kennan and Walker's limited history approach.

nents of income that vary with location, $\mu_{i,j,t}$ and $\theta_{i,j}$, drive migration decisions. Thus, it is possible to ignore the other components in the estimation, as they will drop out of utility. While it is straightforward to include the location- and time-specific mean income, I do not observe the individual's match component. Thus, in estimation I employ the signal extraction of Kennan and Walker.

The signal extraction uses observed wage histories to extract estimates of match components for each individual in each of their observed locations. The central idea of the extraction is that the population-wide distributions can be updated for each individual based on their observed wage histories. For example, if an individual moves and earns an above-average income in both observed locations, it is possible to increase the estimate of the individual's fixed effect (η_i). Likewise, if an individual earns an unexpectedly high income following a move, it is possible to update the estimates of the respective match components ($\theta_{i,j}$).

Isolating the unobserved component of income for each individual in each period and denoting it $\Omega_{i,j(t)}$:

$$\Omega_{i,j(t)} = \theta_{i,j(t)} + \eta_i + e_t$$

it is possible to estimate the population-wide variances of the error (σ^2_θ, σ^2_η, and σ^2_e) by taking sample averages of the elements of the variance-covariance matrix, $\Omega_i\Omega_i'$, where Ω_i is the stacked vector of $\Omega_{i,j(t)}$.[12]

Using these population-wide estimates, Kennan and Walker show that it is possible to update the conditional distribution of the fixed effect for each individual, based on observed income history Ω_i:

$$(\eta_i|\Omega_i) \sim N(\hat{\eta}_i, \hat{\sigma}^2_{i,\eta})$$

[12]The sample average of the diagonal elements will yield an estimate of $\sigma^2_\theta + \sigma^2_\eta + \sigma^2_e$. The sample average of the off-diagonal elements that correspond to the same location will yield an estimate of $\sigma^2_\theta + \sigma^2_\eta$. Finally, the sample average of the off-diagonal elements that correspond to different locations will yield the estimate of σ^2_η, which can be used to separately solve for σ^2_θ, σ^2_η, and σ^2_e.

where:

$$\hat{\eta}_i = \hat{\sigma}^2_{i,\eta}\left[\sum_{\lambda=1}^{\Lambda_i}\frac{\Omega_{i,\lambda}}{\zeta_{i,\lambda}}\right]$$

$$\hat{\sigma}^2_{i,\eta} = \left[\frac{1}{\sigma^2_\eta} + \sum_{\lambda=1}^{\Lambda_i}\frac{\Omega_{i,\lambda}}{\zeta_{i,\lambda}}\right]^{-1}$$

$$\zeta_{i,j} = \sigma^2_\theta + \frac{\sigma^2_e}{S_{i,j}}$$

and where Λ_i represents the number of locations visited by individual i and $S_{i,j}$ represents the number of periods that individual i spends in a given location j.

Using the expected value of individual i's fixed effect by drawing from the above distribution,[13] it is possible to generate a conditional distribution of observed location- and individual-specific match components:

$$(\theta_{i,j}|\Omega_i, E[\eta_i]) \sim N(\hat{\theta}_{i,j}, \hat{\sigma}^2_{i,\theta})$$

where:

$$\hat{\theta}_{i,j} = \sigma^2_{i,\theta}\left[\frac{\Omega_{i,j} - E[\eta_i]}{\pi_{i,j}}\right]$$

$$\hat{\sigma}^2_{i,\theta} = \left[\frac{1}{\sigma^2_\theta} + \frac{1}{\pi_{i,j}}\right]^{-1}$$

$$\pi_{i,j} = \frac{\sigma^2_e}{S_{i,j}}$$

[13]In practice, I simulate using $1,000$ draws for each individual.

Again following Kennan and Walker, I discretize this distribution to three points of support, where the points are fixed over locations, individuals, and time. The following points are shown to be optimal in Kennan (2004): $\mu_j + \Phi^{-1}(\frac{1}{6})$, μ_j, $\mu_j - \Phi^{-1}(\frac{1}{6})$, where $\Phi(\cdot)$ is the standard normal distribution function.

Thus, the signal extraction yields a probability that each individual received a "low," "medium," or "high" match component in each of their observed locations. When considering a new location, an individual expects to receive one of these draws, each with equal probability in expectation, based on the chosen points of discretization.

1.2.2 Moving Costs

In each period, an individual currently located in j chooses a location k from the set of all locations, J. Based on the timing of the model, individuals receive a utility flow associated with their current location j and pay moving costs if their choice of location k is different from j.

The moving cost ($M_{i,j,k}$), depends on the individual's current location (j), choice location (k), prior location (l), and age. In addition to having full information about their current location j, individuals have more information regarding previous locations than locations that they have never visited. In particular, I allow individuals to know the value of the location- and individual-specific match component of income in their prior location. This form of "memory" follows the limited history approach of Kennan and Walker (2006) in allowing individuals to have more information regarding job prospects in places where they have lived previously.

I specify the moving costs associated with choosing location k while currently living in j as:

$$M_{i,j,k} = \psi_0 + \psi_1 distance_{j,k} + \psi_2 distance^2_{j,k} - \psi_3 I_{k \in j.reg} - \psi_4 I_{k=l} + \psi_5 age_i$$

where ψ_0 is a fixed cost of moving, $distance_{j,k}$ is the distance in miles between current location j and chosen location k, $I_{k \in j.reg}$ is an indicator equal to one if k is in the same Census division as j, $I_{k=l}$ is an indicator equal to one if k is equal to individual i's prior location l, and age_i is individual i's current age.

I include both $distance_{j,k}$ and $distance_{j,k}^2$ to allow long-distance moves to be more expensive than more local moves, but with a decreasing effect. Moves that are within the same Census division, or return moves to a prior location, are thought to be less costly. Finally, age is included as it is thought that older individuals (prior to retirement) experience higher costs of moving than younger individuals.[14] Ties to birth location (an implicit form of moving cost) enter into utility directly, as individuals are expected to continually receive utility while residing in the location of their birth.

1.2.3 Utility Specification

Net of moving costs, individuals receive utility from income ($inc_{i,j,t}$), living in their birth location (b_i) or birth region, time-varying location attributes ($z_{j,t}$), fixed observable location attributes (χ_j), fixed unobservable location attributes (ξ_j), and a time-varying idiosyncratic unobservable location attribute ($\epsilon_{i,k,t}$). Thus, flow utility is given by:

$$\begin{aligned} u_{i,j,k,t} + \epsilon_{i,k,t} &= \alpha_{inc} inc_{i,j,t} + z_{j,t}' \alpha_z + \chi_j' \alpha_\chi + \xi_j \\ &\quad + \alpha_{bmsa} I_{j=b_i} + \alpha_{breg} I_{region(j)=region(b_i)} - I_{k \neq j} M_{i,j,k,t} + \epsilon_{i,k,t} \end{aligned}$$

where moving costs enter with an indicator for whether the individual chooses a location other than their current location ($I_{k \neq j}$). The idiosyncratic component of utility ($\epsilon_{i,k,t}$) and the unobservable location attribute (ξ_j) both enter linearly separable.

As the effects of the observable and unobservable fixed location attributes are not separately identified, I collapse all fixed location attributes into a mean flow utility, or

[14] In practice, I use individuals who are aged 21 to 45.

overall quality of life term, δ_j, rewriting utility as:

$$\begin{aligned} u_{i,j,k,t} \ + \ \epsilon_{i,k,t} \ &= \ \alpha_{inc} inc_{i,j,t} \ + \ z'_{j,t}\alpha_z \ + \ \delta_j \\ &\quad + \alpha_{bmsa} I_{j=b_i} \ + \ \alpha_{breg} I_{region(j)=region(b_i)} \ - \ I_{k \neq j} M_{i,j,k,t} \ + \ \epsilon_{i,k,t} \end{aligned}$$

where $\delta_j = \chi'_j \alpha_\chi + \xi_j$.

1.2.4 State Space

The vector of state variables for individual i at time t is denoted $x_{i,t}$. It describes the state of the world at time t and is comprised of all individual characteristics and location attributes that affect utility. The state vector for a given period t includes the individual's current, previous, and birth locations, their age, their current and previous location-specific match components of income, and the current values of location attributes (observed and unobserved) for all locations. The decision variable is equal to the individual's choice at time t and is denoted $d_{i,t}$. Thus, flow utility can be explicitly written as a function of $x_{i,t}$ and $d_{i,t}$:

$$u_t(x_{i,t}, d_{i,t}) \ + \ \epsilon(d_{i,t})$$

I assume that the transition of the state is Markovian, so that $x_{i,t+1}$ depends on $x_{i,t}$ and $d_{i,t}$ only; no additional information is gained by knowing $x_{i,t-1}$. The transition probability of the state vector $x_{i,t}$ is denoted $q(x_{i,t+1}|x_t, d_{i,t})$.

Birth location is fixed and, along with time-invariant location attributes, does not transition. Age, current location, and prior location transition deterministically; in period $t+1$ an agent is two years older, their current location is given by $d_{i,t}$, and, if $d_{i,t}$ necessitates a move, the period t current location will become the period $t+1$ prior location. An individual's location-specific match component of income evolves stochastically in that the value stays fixed if the individual does not move, the value reverts to a known value if the individual returns to their prior location, and, if $d_{i,t}$ involves a move to a new location, the individual gets a draw from the known distribution of match components. Time-varying location attributes evolve stochastically; individuals have beliefs about the distributions from which future amenity values will be drawn. I assume rational expectations, so these distributions are the observed distributions.

The potential size of the state space, or the total number of values state vectors can take on, is the limiting factor in the traditional, full-solution method of estimation. As the number of state variables increases, the size of the state space grows exponentially. The full-solution method, described by Rust (1987), quickly becomes infeasible as the value function needs to be evaluated at every possible combination of the state variables.[15] The two-step estimation method circumvents this problem.

1.2.5 Value Functions

In a world with significant costs of migration, rational agents will not be myopic in their location decisions. Individuals will consider the future stream of utility associated with choosing a new location today, as opposed to the behavior described by a static model where individuals care only about current flow utility. By assuming additive separability of per-period utility over time, it is possible to write lifetime utility, $U(\mathbf{x}, \epsilon, \mathbf{d})$, as the suitably discounted sum of per-period utilities. Thus, individuals choose a decision rule based on:

$$\max_{\mathbf{d}^*=(d_{i,1};\ldots;d_{i,T})} E_{\mathbf{d}}\Big[U(\mathbf{x}, \epsilon, \mathbf{d})\Big] = \max_{\mathbf{d}^*=(d_{i,1};\ldots;d_{i,T})} E_{\mathbf{d}}\left[\sum_{t=1}^{T} \beta^{t-1} \cdot \Big(u_t(x_{i,t}, d_{i,t}) + \epsilon(d_{i,t})\Big)\right]$$

where β is the discount factor.[16]

---

[15] In practice, I take 100 draws for each location in each time period for the five continuous variables – mean wage, crime, housing price, pollution, and population. This implies an approximate state space size of 1.12E+184.

[16] In practice, I set β to 0.9025, or $(0.95)^2$, to account for the biennial nature of the data.

In the lifetime optimization problem current decisions affect both current-period utility and future-periods' utility through the current decision's effect on future states. By assuming (i) the conditional independence of x and ϵ and (ii) that the evolution of the state is Markovian, I am limiting the patterns of dependence in the dynamic process. These, along with the additive separability of flow utility, are the basic assumptions of Rust (1987). This allows me to write the following Bellman equation:

$$V_t(x_{i,t}, \epsilon(d_{i,t})) \quad = \quad \max_{d_{i,t} \in J} [v_t(x_{i,t}, d_{i,t}) + \epsilon(d_{i,t})]$$

where:

$$v_t(x_{i,t}, d_{i,t}) \quad = \quad u_t(x_{i,t}, d_{i,t}) + \beta \int \sum_{x_{i,t+1}} V_{t+1}(x_{i,t+1}, \epsilon(d_{i,t+1}))\; q(x_{i,t+1}|x_{i,t}, d_{i,t})\; dF(\epsilon(d_{i,t+1}))$$

An individual receives flow utility this period, $u_t(x_{i,t}, d_{i,t})$, based on their current state and current location decision, as well as the discounted stream of future payoffs associated with that decision, $\beta \int \sum_{x_{i,t+1}} V_{t+1}(x_{i,t+1}, \epsilon(d_{i,t+1}))\; q(x_{i,t+1}|x_{i,t}, d_{i,t})\; dF(\epsilon(d_{i,t+1}))$. The precise value of this future stream is unknown at time t, as $x_{i,t+1}$ and $\epsilon(d_{i,t+1})$ are only known in expectation.

Assuming that the idiosyncratic error term, $\epsilon(d_{i,t})$, is distributed i.i.d. Type 1 Extreme Value,
$\beta \int \sum_{x_{i,t+1}} V_{t+1}(x_{i,t+1}, \epsilon(d_{i,t+1}))\; q(x_{i,t+1}|x_{i,t}, d_{i,t})\; dF(\epsilon(d_{i,t+1}))$ can be replaced with the familiar Logit inclusive value:

$$v_t(x_{i,t}, d_{i,t}) \quad = \quad u_t(x_{i,t}, d_{i,t}) + \beta \sum_{x_{i,t+1}} ln \left[\sum_{j=1}^{J} exp \left(v_{t+1}(x_{i,t+1}, d_{i,t+1} = j) \right) \right] q(x_{i,t+1}|x_{i,t}, d_{i,t})$$

This is the main estimating equation of the model. The computational burden of the traditional solution method comes from the recursive nature of this equation; there is a v_t on the left-hand side and v_{t+1} on the right-hand side. Specifically, there are J number of v_{t+1} terms on the right-hand side, which describe the value of choosing any of the J locations in period $t+1$. This is where the insights of Hotz and Miller (1993) and Arcidiacono and Miller (2007) greatly simplify the problem and allow for a two-step estimation routine.

Algebraically, it is possible to multiply and divide the inclusive value term by the value of choosing a particular location h in period $t+1$, given that $d_{i,t} = k$:

$$\begin{aligned} v_t(x_{i,t}, d_{i,t} = k) &= u_t(x_{i,t}, d_{i,t} = k) \\ +\beta \sum_{x_{i,t+1}} ln\Bigg[\sum_{j=1}^{J} exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = j)\Big) & \frac{exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big)}{exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big)}\Bigg] q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \end{aligned}$$

Separating out the $exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big)$ in the numerator and dividing through by the $exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big)$ in the denominator yields:

$$\begin{aligned} v_t(x_{i,t}, d_{i,t} = k) &= u_t(x_{i,t}, d_{i,t} = k) \\ +\beta \sum_{x_{i,t+1}} ln\Bigg[\sum_{j=1}^{J} exp\Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = j) - v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big)\Bigg] & q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\ + \beta \sum_{x_{i,t+1}} \Big[v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big] & q(i, x_{t+1}|x_{i,t}, d_{i,t} = k) \end{aligned}$$

The third right-hand side term in the above equation,

$$\beta \sum_{x_{i,t+1}} \Big[v_{t+1}(x_{i,t+1}, d_{i,t+1} = h) \Big] q(x_{i,t+1} | x_{i,t}, d_{i,t} = k) \ ,$$

or the expected value of choosing location h in period $t+1$, can be written as the sum of a period $t+1$ flow utility and the associated continuation value. Similar to the previous normalization, it is possible to normalize this continuation value relative to choosing some location g in period $t+2$:

$$\begin{aligned}
v_t(x_{i,t}, d_{i,t} = k) \ &= \ u_t(x_{i,t}, d_{i,t} = k) \\
&+ \ \beta \sum_{x_{i,t+1}} ln \Bigg[\sum_{j=1}^{J} exp \Big(v_{t+1}(x_{i,t+1}, d_{i,t+1} = j) - v_{t+1}(x_{i,t+1}, d_{i,t+1} = h) \Big) \Bigg] q(x_{i,t+1} | x_{i,t}, d_{i,t} = k) \\
&+ \ \beta \sum_{x_{i,t+1}} \Big[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h) \Big] q(x_{i,t+1} | x_{i,t}, d_{i,t} = k) \\
&+ \ \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln \Bigg[\sum_{j=1}^{J} exp \Big(v_{t+2}(x_{i,t+2}, d_{i,t+2} = j) - v_{t+2}(x_{i,t+2}, d_{i,t+2} = g) \Big) \Bigg] \\
&\quad \cdot q(x_{i,t+2} | x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1} | x_{i,t}, d_{i,t} = k) \\
&+ \ \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \Big[v_{t+2}(x_{i,t+2}, d_{i,t+2} = g) \Big] q(x_{i,t+2} | x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1} | x_{i,t}, d_{i,t} = k)
\end{aligned}$$

Finally, it is possible to again expand the final right-hand side value function into a flow utility and a continuation value, and normalize the continuation value by the value of choosing some location m in period $t+3$:

$$
\begin{aligned}
v_t(x_{i,t}, d_{i,t} = k) \;\; &= \;\; u_t(x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta \sum_{x_{i,t+1}} ln\Bigg[\sum_{j=1}^{J} exp\Bigg(v_{t+1}(x_{i,t+1}, d_{i,t+1} = j) - v_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Bigg)\Bigg] q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta \sum_{x_{i,t+1}} \Bigg[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Bigg] q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln\Bigg[\sum_{j=1}^{J} exp\Bigg(v_{t+2}(x_{i,t+2}, d_{i,t+2} = j) - v_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\Bigg)\Bigg] \\
&\quad \cdot q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \Bigg[u_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\Bigg] \\
&\quad \cdot q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} ln\Bigg[\sum_{j=1}^{J} exp\Bigg(v_{t+3}(x_{i,t+3}, d_{i,t+3} = j) - v_{t+3}(x_{i,t+3}, d_{i,t+3} = m)\Bigg)\Bigg] \\
&\quad \cdot q(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g) q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\; \beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} \Bigg[v_{t+3}(x_{i,t+3}, d_{i,t+3} = m)\Bigg] \\
&\quad \cdot q(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g) q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k)
\end{aligned}
$$

While these expansions appear to have complicated the estimation equation, they actually allow for an important simplification. Consider the probability of choosing a particular location c in a given time τ. Given the Logit framework, this probability can be written as:

$$
\begin{aligned}
Pr(d_{i,\tau} = c | x_{i,\tau}) &= \frac{exp\Big(v_\tau(x_{i,\tau}, d_{i,\tau} = c)\Big)}{\sum_{j=1}^{J} exp\Big(v_\tau(x_{i,\tau}, d_{i,\tau} = j)\Big)} \\
&= \frac{1}{\sum_{j=1}^{J} exp\Big(v_\tau(x_{i,\tau}, d_{i,\tau} = j) - v_\tau(x_{i,\tau}, d_{i,\tau} = c)\Big)}
\end{aligned}
$$

or as:

$$
Pr(d_{i,\tau} = c | x_{i,\tau})^{-1} = \sum_{j=1}^{J} exp\Big(v_\tau(x_{i,\tau}, d_{i,\tau} = j) - v_\tau(x_{i,\tau}, d_{i,\tau} = c)\Big)
$$

The right-hand side of the above equation is equivalent to the right-hand side normalized continuation values in the choice-specific value function. Thus, it is possible to write the choice-specific value function as:

$$
\begin{aligned}
v_t(x_{i,t}, d_{i,t} = k) \;\; &= \;\; u_t(x_{i,t}, d_{i,t} = k) \\
&+ \;\beta \sum_{x_{i,t+1}} ln\Big[Pr(d_{i,t+1} = h|x_{i,t+1})^{-1}\Big] q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\beta \sum_{x_{i,t+1}} \Big[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\Big] q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln\Big[Pr(d_{i,t+2} = g|x_{i,t+2})^{-1}\Big] \\
&\quad \cdot q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \Big[u_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\Big] \\
&\quad \cdot q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} ln\Big[Pr(d_{i,t+3} = m|x_{i,t+3})^{-1}\Big] \\
&\quad \cdot q(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g) q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \;\beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} \Big[v_{t+3}(x_{i,t+3}, d_{i,t+3} = m)\Big] \\
&\quad \cdot q(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g) q(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h) q(x_{i,t+1}|x_{i,t}, d_{i,t} = k)
\end{aligned}
$$

These future conditional choice probabilities, $Pr(d_{i,t}|x_{i,t})$, are estimated in a separate first-step. They can therefore be thought of as data in the above equation and replaced by $\widehat{Pr}(d_{i,t}|x_{i,t})$. The transition probabilities, $q(x_{i,t+1}|x_{i,t}, d_{i,t})$, can be similarly recovered in a preliminary procedure and replaced by $\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t})$:

$$
\begin{aligned}
v_t(x_{i,t}, d_{i,t} = k) \;\; &= \;\; u_t(x_{i,t}, d_{i,t} = k) \\
&+ \; \beta \sum_{x_{i,t+1}} ln\left[\widehat{Pr}(d_{i,t+1} = h|x_{i,t+1})^{-1}\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \; \beta \sum_{x_{i,t+1}} \left[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln\left[\widehat{Pr}(d_{i,t+2} = g|x_{i,t+2})^{-1}\right] \\
&\quad \cdot \widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \left[u_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\right]\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \; \beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} ln\left[\widehat{Pr}(d_{i,t+3} = m|x_{i,t+3})^{-1}\right] \\
&\quad \cdot \widehat{q}(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g)\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&+ \; \beta^3 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \sum_{x_{i,t+3}} \left[v_{t+3}(x_{i,t+3}, d_{i,t+3} = m)\right] \\
&\quad \cdot \widehat{q}(x_{i,t+3}|x_{i,t+2}, d_{i,t+2} = g)\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k)
\end{aligned}
$$

Now, the main estimating equation has been re-written to eliminate $J-1$ of the original J right-hand side value functions. Exploiting the “finite dependence” simplification of Arcidiacono and Miller will eliminate the remaining one. Before describing this final step, however, it is useful to consider the economic meaning of the remaining right-hand side value function ($v_{t+3}(x_{i,t+3}, d_{i,t+3} = m)$) – *i.e.*, the value of choosing location m in period $t+3$, when the individual's current ($t+2$ choice) location is g and prior ($t+1$ choice) location is h. The fact that the agent chose location k in period t does not affect the period

$t+3$ decision, as "memory" only extends to the prior location. In this case, the dependence on initial choice can be broken after two periods, while in a model with no "memory," dependence could be broken after only one period.[17]

Using the fact that the location parameter in a Logit model is not identified, *i.e.*, that only relative values matter, it is possible to difference the above value function with respect to the value function associated with another choice (*i.e.*, estimate $(v_t(x_{i,t}, d_{i,t} = k) - v_t(x_{i,t}, d_{i,t} = c))$, where the future component of $v_t(x_{i,t}, d_{i,t} = c)$ has been expanded in the manner just described). With dependence on initial choice broken, the final right-hand side term has the same value in both cases. In addition, as the value of choosing and particular location in period $t + 3$ is independent of the period t choice, the probability of choosing a particular location m in period $t + 3$ is also independent of the period t choice. Thus, the final two right-hand side terms will drop out when differenced. In general, the model must be expanded to one period further than "memory" allows; if an agent can "remember" current location and prior location, the value function must be extended three periods, as above.

The estimating equation therefore becomes:

[17]This would allow the methodology to be used on publicly available IPUMS data, which provides information on the individual's birth location and two other observed locations (*e.g.*, location in 1985 and in 1990). The latter information can be used to model one migration decision. The current specification requires data on birth location and at least three other observed locations, or two migration decisions.

$$
\begin{aligned}
v_t(x_{i,t}, d_{i,t} = k) - v_t(x_{i,t}, d_{i,t} = c) \;\; &= \;\; u_t(x_{i,t}, d_{i,t} = k) - u_t(x_{i,t}, d_{i,t} = c) \\
&+ \; \beta \sum_{x_{i,t+1}} ln\left[\widehat{Pr}(d_{i,t+1} = h|x_{i,t+1})^{-1}\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&- \; \beta \sum_{x_{i,t+1}} ln\left[\widehat{Pr}(d_{i,t+1} = h|x_{i,t+1})^{-1}\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = c) \\
&+ \; \beta \sum_{x_{i,t+1}} \left[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&- \; \beta \sum_{x_{i,t+1}} \left[u_{t+1}(x_{i,t+1}, d_{i,t+1} = h)\right]\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = c) \\
&+ \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln\left[\widehat{Pr}(d_{i,t+2} = g|x_{i,t+2})^{-1}\right]\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&- \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} ln\left[\widehat{Pr}(d_{i,t+2} = g|x_{i,t+2})^{-1}\right]\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = c) \\
&+ \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \left[u_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\right]\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = k) \\
&- \; \beta^2 \sum_{x_{i,t+1}} \sum_{x_{i,t+2}} \left[u_{t+2}(x_{i,t+2}, d_{i,t+2} = g)\right]\widehat{q}(x_{i,t+2}|x_{i,t+1}, d_{i,t+1} = h)\widehat{q}(x_{i,t+1}|x_{i,t}, d_{i,t} = c)
\end{aligned}
$$

While it is still quite complicated, note that this equation is linear in the structural parameters. These parameters can therefore be estimated using a straight-forward Logit procedure without recursion, after integrating out over the transition of the state. The estimation of these value functions will be discussed in more detail in Section 1.4.

1.3 Data

1.3.1 Choice Set

As the spatial determinants of national migration decisions are the focus of my research, the geographic definition of the choice set is critical; if the locations are chosen on too fine a level, there will be few observed moves to and from each location; however, if the locations are chosen on too broad a level (such as states or regions), it becomes impractical to define spatially varying amenities or to infer preferences for them. In a national model of migration, metropolitan areas seem to be a natural unit of geography. However, with standard panel datasets, sparse data becomes an issue when including all 362 recognized metropolitan areas.[18]

Although my methodology could handle a much larger choice set, I use the 50 largest metropolitan statistical areas (MSAs) in the contiguous United States, as represented in my primary data source, the National Longitudinal Survey of Youth 1979 (NLSY79).[19] To limit attrition from the panel, I additionally include 9 "catch-all" locations, defined by non-MSA Census divisions. Thus, the final choice set is comprised of 59 locations. The 50 MSAs represent approximately 39% of the U.S. population in 2000 and approximately 60% of my NLSY79 sample. The geographic distribution of these locations can be seen in Figure 1.1.

1.3.2 Individual Data

The primary dataset I employ is the National Longitudinal Survey of Youth, 1979 panel. This panel dataset follows over 12,000 individuals aged 14 to 22 in the first round of interviews in 1979. Beginning in 1994, the Bureau of Labor Statistics switched from

---

[18]In future work, the Longitudinal Employer - Household Dynamic (LEHD) panel dataset, which contains basic demographic, income, and detailed geographic data for millions of individuals based on Unemployment Insurance records, could be used to estimate a model with all 362 MSAs.

[19]While the largest MSAs in the NLSY include the most recognized cities in the U.S. (*e.g.*, New York, Chicago, Los Angeles), there are a number of smaller cities that contain a large number of NLSY respondents relative to their size – *e.g.*, Hickory-Morgantown-Lenoir, NC, Lima, OH, Great Falls, MT.

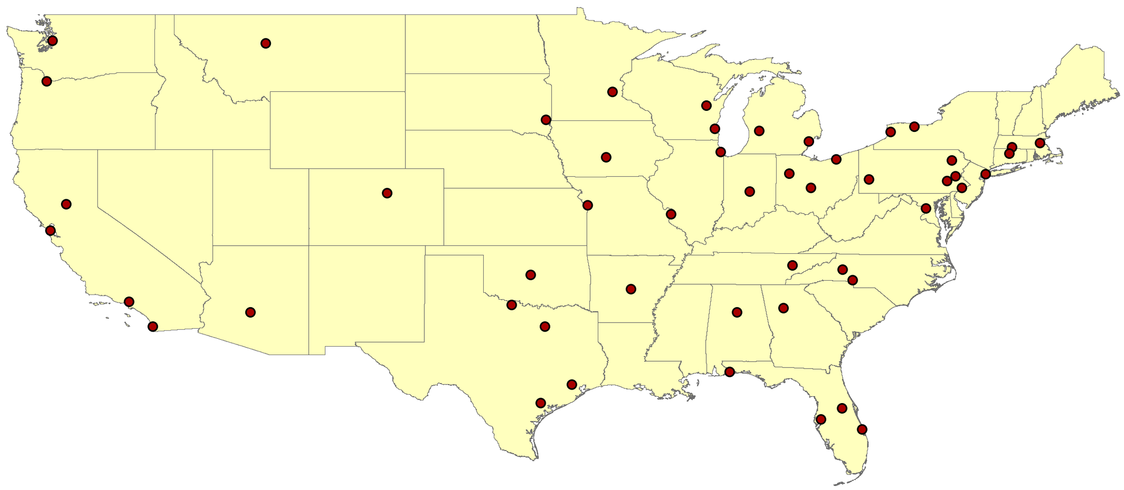

Figure 1.1: Metropolitan Areas of the Choice Set

an annual to a biennial interview basis. I use biennial interviews from 1986 to 2004, with individuals aged 21 to 29 in 1986 and 39 to 45 in 2004 (the panel is unbalanced). The restricted-access geocode supplement of the NLSY79 provides detailed information (county-level) on the individual's geographic location, including birth location, in addition to the detailed demographic data given by the public-access files. I aggregate this county-level data to the level of the MSA, as defined by the U.S. Census Bureau in 1990.

Following Kennan and Walker, I cut the sample to include a homogenous sample of high school educated white males, with no military or college experience.[20] In addition, I drop individuals who were born outside of the contiguous United States. Finally, I drop observations following a missed interview by the individual. This leaves me with a final sample of 1,123 individuals with 7,176 observations and 478 observed moves between locations in the choice set.[21]

In particular, I use information on individuals' birth locations and migration decisions, age, AFQT score,[22] and income (defined as the sum of wage and business income).

Summary statistics are given in Table 1.1 (the sample is comprised of 7,176 person-year observations with 1,123 distinct individuals).

Table 1.1: Summary Statistics for the NLSY79 Sample

Variable	Obs.	Mean	Std. Dev.
Age	7176	31.88	5.63
Income (2000 dollars)	7176	32,997.26	22,572.47
Live in MSA of Birth	7176	0.68	0.46
Number of Moves	1123	0.49	0.93
Number of Interviews	1123	7.93	2.18
AFQT	1123	54.08	25.91

Finally, in the estimation of location-specific incomes, I supplement the NLSY79 sample with a sample of Current Population Survey (CPS) respondents. This sample is also trimmed to white, high school educated males with no college or military experience. In

[20]One may worry that if college attendance rates were changing over the period, differing cohort effects would limit the homogeneity of the panel. However, national college attendance rates for 25 to 29 year old white males remained fairly constant over the age-relevant period of 1982 to 1993. Please see the U.S. Health and Human Services Department report findings at http://aspe.hhs.gov/hsp/97trends/ea1-6.htm for more detail.

[21]Although this is a relatively small number of moves compared to static analyses, which use cross-sectional data, it is more than twice the number of moves observed by Kennan and Walker (213). Future access to richer panel datasets will increase this number significantly.

[22]The Armed Forces Qualifying Test is an general aptitude test given to all respondents of the NLSY79.

addition, I trim to U.S.-born individuals of the relevant age cohort.[23]

1.3.3 Location Attributes

As any time-invariant characteristics would be absorbed by the location-specific fixed effect, δ_j, I collect a panel of annual time-varying data comprised of crime, median housing price, air pollution, and population variables for each MSA over the relevant period. I construct a matrix of distances using the "Great Circle" algorithm[24] with geographic coordinate data taken from the U.S. Census Bureau.

Crime

The crime data I employ are taken from the Federal Bureau of Investigation's annual report entitled "Crime in the United States." These annual reports include the total reported violent and property crime incidents,[25] as reported by over 17,000 law enforcement agencies across the United States. These data are given at the city-, MSA-, state-, region-, and national- level on an annal basis beginning in 1930.

Following Savageau and D'Agostino (2000), I compute a measure of crime that collapses the violent crimes and property crimes into a single index for each MSA in each year. As it is assumed that individuals respond to violent crimes more than property crimes, the index weights violent crimes ten times as heavily. I use a slightly modified version of the index by calculating it in per-capita terms.[26] Specifically, I use:

[23]The five MSAs of Fort Pierce-Port St. Lucie, FL, Great Falls, MT, Lima, OH, Victoria, TX, and Wichita Falls, TX do not have full CPS coverage over the relevant sample period. In these cases, I include individuals residing within the larger Public Use Microdata Area (PUMA).

[24]This is also known as the Haversine formula (Sinnott, 1984).

[25]Property crimes include burglary, larceny-theft, and motor vehicle theft. Violent crimes include murder, manslaughter, forcible rape, robbery, and aggravated assault. As not all reporting agencies collect data on rape, I omit it from my aggregate measure of violent crime.

[26]I use the FBI's measure of population here to ensure a proper per-capita measure.

$$crime_{j,t} = \frac{violent_{j,t} + \frac{property_{j,t}}{10}}{population_{j,t}}$$

Median House Price

I include a measure of median house price to control for differences in cost of living both across MSAs and across time periods. Using data from the Office of Federal Housing Enterprise Oversight's (OFHEO) Housing Price Index (HPI) and estimates from the National Association of Realtors, I construct a panel of median house prices. The HPI, which sets 1995 as a base year for each MSA, reports a weighted, repeat-sales index for single-family homes within each MSA going back as far as 1975. I translate these appreciation figures into annual median house prices using a 2004 cross-section of median single-family house prices provided by the National Association of Realtors. Finally, all prices are converted to 2000 dollars using the Consumer Price Index.

Air Pollution

The measure of air pollution I use is the ambient concentration of particulate matter. The Environmental Protection Agency provides annual emissions figures in their National Emissions Inventory from nearly $6,000$ different sources. The data I use are county-level estimates of particulate matter concentration (PM10) generated from source data on total particulates and sulfur dioxide, a PM10 precursor, using the source to county receptor matrix of the Climatological Regional Dispersion Model.[27] I compute annual MSA-level pollution estimates by aggregating the county-level data to the 1990 geographical definition of MSA boundaries.

[27] The county-level dataset was created by Nat Keohane who generously provided it for this research.

Population

Finally, I collect annual county-level population estimates from the U.S. Census Bureau. I compute annual MSA-level population estimates by aggregating the county-level data to the 1990 geographical definition of MSA boundaries. Thus, this figure captures changes in population density without the confounding effects of changes to the MSA boundaries.[28]

Summary statistics of the location attributes are presented in Table 1.2.

Table 1.2: Summary Statistics for the Location Attributes

Variable	Obs.	Mean	Std. Dev.
Crime (per capita)	50	0.0116	0.0046
Median House Price (hundred thousand 2000 dollars)	50	130.55	64.58
Pollution (PM10 in $\mu g/m^3$)	50	27.79	7.60
Population (millions)	50	2.02	2.19
New England	50	0.07	0.25
Middle Atlantic	50	0.15	0.36
East North Central	50	0.17	0.38
West North Central	50	0.10	0.30
South Atlantic	50	0.15	0.36
East South Central	50	0.05	0.22
West South Central	50	0.12	0.33
Mountain	50	0.07	0.25
Pacific	50	0.12	0.33

1.4 Estimation

The computational burden of the model is greatly reduced by the two-step estimation strategy described in Arcidiacono and Miller (2007). The first step includes all of the estimation procedures that are performed outside of the dynamic routine. In particular, I estimate incomes, transition probabilities of the variables that evolve stochastically, and conditional choice probabilities. In the second step, I estimate the remaining structural

[28]The geographic boundaries of some U.S. counties changed over the period of my sample. However, these changes did not affect the definition of any of the fifty chosen MSAs.

parameters of the utility function, taking the first-stage estimates as given.

1.4.1 First Step

Income

In the estimation of location-specific mean income levels, I supplement the NLSY79 sample with data taken from the CPS over the relevant periods, locations, and cohort. Using the CPS, I regress reported income on a set of age and MSA dummies, yielding estimates of $f(age_{i,t})$ and $\mu_{j,t}$:

$$inc^{CPS}_{i,j,t} \;=\; f(age_{i,t}) \;+\; \mu_{j,t} + \epsilon^{CPS}_{i,j,t}$$

I am then able to take these estimates to the NLSY79 sample, where I regress the deviations from predicted income on individual characteristics:[29]

$$inc_{i,j,t} \;-\; \hat{f}(age_{i,t}) \;-\; \hat{\mu}_{j,t} \;=\; \omega_i'\gamma \;+\; (\theta_{i,j} + \eta_i + e_t)$$

allowing me to isolate the unobserved, idiosyncratic portion of income:

$$inc_{i,j,t} \;-\; \hat{f}(age_{i,t}) \;-\; \hat{\mu}_{j,t} \;-\; \omega_i'\hat{\gamma} \quad=\quad \theta_{i,j} + \eta_i + e_t \quad=\quad \Omega_{i,j(t)}$$

I find population-wide estimates of standard deviations for $\theta_{i,j}$, η_i, and e_t to be (in 2000

[29]As the sample is cut to a homogenous sample, I follow Kennan and Walker in using AFQT score as the sole individual characteristic.

dollars):[30]

$$\hat{\sigma}_\eta = \$7,073.40$$

$$\hat{\sigma}_\theta = \$15,499.21$$

$$\hat{\sigma}_e = \$13,869.55$$

$$\hat{\sigma}_{inc} = \$21,968.68$$

and the estimated points of support of the distribution of match components in each location to be (also in 2000 dollars):

$$\mu_{j,t} - \$8,775.40$$

$$\mu_{j,t}$$

$$\mu_{j,t} + \$8,775.40$$

where $\mu_{j,t}$ is the mean income in location j in period t.

Transition Probabilities

Also in the first stage, I estimate the transitions for each of the five stochastic variables of the model (mean income, crime rate, housing price, pollution level, and annual population) by assuming AR-1 processes. To maximize use of the data, I pool observations across locations and regress the current value of these variables on a constant term, the lagged value of the variable, a set of Census division dummies, and a set of dummies describing the

[30]These estimates imply that approximately 41% of the total variance of earnings can be attributed to the transitory component of income. This figure is larger than that estimated in Kennan and Walker (33%). However, both are in line with the previous research of Gottschalk and Moffitt (1994).

MSA's quintile of 1990 population. This yields a set of predicted values for each variable in each period. The residuals from these regressions are used to define the distributions from which each expected value is drawn.

Conditional Choice Probabilities

Finally, the conditional choice probabilities are recovered before proceeding to the dynamic estimation problem. In an ideal world, a first-stage estimate of these probabilities would be found using a fully non-parametric method, such as a "bin" estimator, where the probability of choosing a particular location could be estimated as the fraction of individuals who choose that location (conditional on being at a particular point in the state space). However, these probabilities need to be calculated at each and every possible state of the world, with this number approaching infinity in the current specification. Thus, I estimate a reduced-form approximation of the conditional choice probabilities, specifying a flexible functional form. Within the Logit framework, I include a sizeable number of higher-order and interaction terms, estimating over 40 parameters in addition to a set of location fixed-effects, which enter linearly.[31]

1.4.2 Second Step

The first step yields transitions of the state variables and the conditional probabilities of choosing any particular location, given any particular state. These estimates are used in the second-step, where the remaining structural parameters of the specified utility function are estimated.

Using the insights of Arcidiacono and Miller (described in Section 1.2.5), the second step is estimated as a linear-in-parameters Logit. As the location fixed-effects (the δ's) enter linearly, it is possible to employ a Berry (1994) contraction mapping to estimate them,

[31] I estimate the fixed effects using a Berry (1994) contraction mapping and the remaining parameters using Maximum Likelihood.

further limiting the number of parameters over which I need to search in a Maximum Likelihood routine. The remaining parameters (the α's), are estimated by maximizing the following log-likelihood:

$$L(\alpha) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{J} log\Big(\frac{exp(v_t(x_{i,t}, d_{i,t} = k))}{\sum_{j=1}^{J} exp(v_t(x_{i,t}, d_{i,t} = j))}\Big) \cdot I_{[d_{i,t}=k]}$$

where $I_{[d_{i,t}=k]}$ is an indicator equal to one if individual i chooses location k in period t.

The state vector is updated in each period using the estimated transitions from the first step. In practice, I integrate out over the distributions of the stochastic variables by taking 100 draws from their distributions, which are generated using the residuals from the AR-1 regressions. The second-stage algorithm takes approximately six hours to converge, given reasonable starting values.

Finally, a willingness to pay for a marginal change in a particular attribute can be easily calculated as the coefficient on the attribute divided by the coefficient on income, as the variables enter linearly.[32]

1.5 Results

Table 1.3 reports the parameter estimates from the second-stage dynamic estimation. Estimates have the predicted sign and most coefficients are significant at the 5% level. Income and housing price are given in tens of thousands of 2000 dollars, population is reported in millions of individuals, distance is measured in thousands of miles, and pollution is given in $\frac{\mu g/m^3}{100}$

These results imply a willingness to pay to avoid a one unit (one $\mu g/m^3$) increase in air pollution of \$198.22 in 2000 dollars. Reported in 1982-1984 dollars, this figure is \$114.96.

[32] Considering larger changes, one would also need to account for the effects on expected future flow utilities in the calculation of compensating income variations. In this paper, I focus only on willingness to pay for marginal changes in air pollution, dealing with large changes in ongoing work.

Table 1.3: Results from Dynamic model

Log Likelihood	-3097.1626	
	Coefficient	t-statistic
Moving Cost - Fixed Cost	-3.8971	-40.8531
Moving Cost - Distance	-1.9565	-11.6354
Moving Cost - Distance^2	0.3908	5.1235
Moving Cost - Same Region	0.8588	12.2621
Moving Cost - Age	-0.1019	-24.9054
Moving Cost - Last MSA	0.2111	2.0557
Income	0.1408	3.9983
Living in MSA of Birth	0.3550	18.4645
Living in Region of Birth	0.0001	0.0038
Housing Price	-0.0337	-0.0056
Crime	-0.1853	-0.0389
Pollution	-0.2791	-1.9906
Population	-1.3334	-2.3551

In a static hedonic analysis, Chay and Greenstone (2005) find a comparable willingness to pay of approximately \$22.00 (also in 1982-1984 dollars). However, their model ignores both forward-looking behavior and the costs associated with migration. Each of these omissions could lead to downward-biased estimates.

It is interesting to note that the coefficient on crime is not significant; this is not surprising, as crime is often extremely localized within an MSA and individuals can choose to avoid it (to a certain degree) based on neighborhood choice. However, this model could be used to derive a willingness to pay for any time-varying attribute, such as crime, in the same manner as pollution. The willingness to pay for a fixed attribute could be similarly estimated by decomposing the location specific fixed effects into observable and unobservable components via OLS regression:

$$\delta_j \;=\; \chi_j'\alpha_\chi \;+\; \xi_j$$

I am unable to do so in the current application only because of the limitations imposed by the current choice set size (*i.e.*, $J = 59$) that results from using NLSY data.

Finally, the moving costs suggested by the above coefficients are very large. The constant, fixed cost of moving implies a dollar value of $\$277,857.14$. The cost of a one-mile move by a twenty-year old individual would be $\$335,714.30$. Although large, this is in line with the previous literature. The comparable figure (converted to 2000 dollars) in Kennan and Walker's analysis is a slightly larger $\$363,085.80$. As discussed in Kennan and Walker, one would expect this cost estimate to be large as it captures the cost an individual would expect to face if forced to move to an arbitrary location in an arbitrary time period. By allowing an individual to choose the best location, but still conditional on being forced to move in an arbitrary time period, would reduce costs by $128,211.[33]

To further examine the consequences of ignoring either dynamics or psychological ties to certain locations, I estimate a static specification and a dynamic specification that ignores ties to birth location using the NLSY79 sample. Results from these specifications are presented in Tables 1.4 and 1.5 respectively.

Table 1.4: Results from Static model

Log Likelihood	-3118.0622	
	Coefficient	t-statistic
Moving Cost - Fixed Cost	-1.0863	-28.7229
Moving Cost - Distance	-2.0204	-12.0548
Moving Cost - Distance^2	0.5622	7.3940
Moving Cost - Same Region	0.6562	9.4096
Moving Cost - Age	-0.0927	-20.325
Moving Cost - Last MSA	-0.9613	-4.4994
Income	0.3167	5.1757
Living in MSA of Birth	3.0761	84.2343
Living in Region of Birth	0.0023	0.0354
Housing Price	-0.0017	-0.2552
Crime	-0.9525	-0.1378
Pollution	-0.2401	-1.8793
Population	-1.9054	-1.5985

[33] $128,211 is calculated as $log(J-1)/\alpha_{inc}$.

In the static specification, agents choose a location based on the current level of attributes. Thus, the value function is comprised of flow utility and a choice-specific shock only; agents ignore the continuation value associated with current choice. Note that the static specification does, however, still allow individuals to have psychological ties to their birth locations.

The results from the static model imply that individuals sort on the basis of both current and expected future levels of pollution (in particular, choosing locations where pollution levels are expected to fall). This leads to estimates of willingness to pay that are downward-biased in the static framework. In particular, this model yields a willingness to pay for air quality that is 62% lower than in the dynamic specification. Specifically, this model finds a willingness to pay to avoid a one unit increase in air pollution of approximately $43.55 in 1982-1984 dollars.

Finally, it is also interesting to note that in the static model, the sign on "Last MSA" has flipped; without dynamic concerns, the importance of a prior location to agents is unclear, as opposed to the dynamic model, where agents consider the possibility of future return migration.

Table 1.5: Results from Dynamic model with no ties to Birth Location

Log Likelihood	-3127.3720	
	Coefficient	t-statistic
Moving Cost - Fixed Cost	-4.1609	-0.6533
Moving Cost - Distance	-1.9292	-0.4152
Moving Cost - Distance^2	0.3995	0.0298
Moving Cost - Same Region	0.9616	0.2322
Moving Cost - Age	-0.1038	0.0066
Moving Cost - Last MSA	0.0020	0.0002
Income	0.3781	0.0041
Housing Price	-0.0337	-0.0056
Crime	0.7036	3.3274
Pollution	-0.3355	-0.1188
Population	-4.8837	-2.6464

In the next specification, agents are both forward-looking and experience costs associated with migration in the traditional sense (*i.e.*, there is a fixed cost moving and costs increase with age and distance between locations). However, often over-looked are the emotional ties to one's "home," or birth location. This specification ignores individual's preference for living in the MSA or region of their birth. Bayer, Keohane, and Timmins (2006) show using 2000 Census data, that more than sixty percent of household heads are currently living in the Census division of their birth, with a higher number for high-school graduates, such as those used in this application.

If individuals have ties to locations with relatively high levels of pollution because of birth, a naïve model that ignores these ties would, by omission, interpret them as a preference for birth location attributes. This would lead estimates of willingness to pay for air quality to be biased downwards. There is a positive correlation between birth location and pollution (the correlation coefficient is 0.22, measured in the first year of my panel, 1986), implying that this source of bias may be a concern.

The results presented in Table 1.5 imply that individuals have ties to birth locations with relatively worsening air quality, as identification comes off of variation through time. These willingness to pay results are again substantially lower than in the base model; $52.21 versus $114.96 in 1982-1984 dollars. In addition, the omission of these ties leads many of the other coefficients to take on the wrong signs.

1.6 Extensions

The goal of the preceding exercise was to illustrate the important role played by dynamics (both costly migration and forward-looking behavior) in non-market valuation. In an application to air quality, the impacts are substantial. However, since I have recovered the parameters of the utility function that determine migration behavior, I can also use the estimated model to perform counterfactual simulations of individual migration decisions. Of particular interest is the role of amenities in the migration patterns observed over the last twenty years in the United States. Isolating the relative weights on income prospects and on non-pecuniary amenity values in migration trends, I find in preliminary simulations that amenities play a significant role (as opposed to falling wages in declining industries alone) in the recent flows out of Northeast and Midwest and into the South, Southwest, and West of the United States. Driving this finding is the result that income and amenity values are positively correlated in metropolitan areas.

In addition, these simulations can be used to examine the ex post distributions of the individual- and location-specific match components of income. Following the basic insights of Roy (1951), observed income distributions will have higher means than offered income distributions, as individuals choose locations that offer the highest wages. Although I capture only the partial equilibrium effects while holding location-specific mean incomes fixed, preliminary results show that the ex post distribution of location- and individual-specific match components has both a larger mean and a larger variance in the counterfactual world in which amenities do not enter utility. This implies that inter-location convergence of incomes could be in part explained by geographic differences in amenities.[34]

Finally, the estimates presented in this paper apply to the homogenous sample of white, high-school educated males. I plan to estimate the model using different cohorts from the NLSY79. The nature of my model will then allow me to shed light on the determinants of observed wage gap dynamics based on individual characteristics, such as education and race. By failing to account for expected wage profiles, as well as the effects of Roy sorting, simple static models could give biased estimates of differences in offered income distributions. For example, if college educated individuals are more likely to locate based on higher expected growth in wages, estimates of the college wage premium will be biased downwards. Alternatively, if college educated individuals have lower moving costs, they will be more

[34]Barro and Sala-i-Martin (1992) find convergence (although less than the neoclassical model would predict) across the contiguous U.S. since 1840.)

likely to move and get a high match component; this type of Roy (1951) sorting will cause estimates of offered wage gap to be biased upwards.[35] As my model accounts for expected wage profiles and sorting, estimating it separately for different demographic groups would allow me to decompose the wage gaps and explain what portion is determined by true differences in offered wages versus the effects of sorting and dynamics. Similarly, I could examine differences in wages across cities based on city characteristics, such as city size.

1.7 Conclusion

In this paper, I estimate a dynamic model of location choice where the choice set is defined at the level of the metropolitan area. I employ the two-step, Hotz and Miller -style estimation routine recently developed in Arcidiacono and Miller (2007), which permits me to include a rich descriptive set of local attributes, including a set of location fixed effects. The inclusion of these amenity values not only allows for a more realistic specification of individual utility, but also allows for the use of a dynamic migration model in a non-market valuation exercise.

In the first step of the estimation routine, I flexibly estimate reduced-form conditional choice probabilities, transition probabilities of the state variables, and individual incomes. In the second step, I estimate the structural parameters of individuals' utility functions. The overall computational burden of the estimator is low, with the second step estimated as a linear-in-parameters Logit. I estimate the model using a homogenous sample of white, high-school educated males from the NLSY79 panel, over the period 1986 to 2004.

My particular application recovers the willingness to pay for clean air. Using a panel of annual air pollution data for each metropolitan area, I find the willingness to pay (in 2000 dollars) to be \$198.22 for a one unit decrease in annual pollution (a one $\mu g/m^3$ decrease in

[35] Dahl (2002) and Bayer, Khan, and Timmins (2006) outline two different approaches for controlling for sorting when recovering estimates of the offered wage distribution from observed accepted wage distributions. However, neither paper controls for dynamics.

particulate matter). When calculated at mean income and mean pollution, this corresponds to a willingness to pay elasticity with respect to air quality of approximately 0.17. I find evidence of a downward bias in estimates produced using either a static sorting model or a dynamic model that ignores the emotional ties to individuals' birth locations. This indicates that individuals are both forward-looking with respect to location amenities and have ties to relatively polluted areas.

In future work, I plan to both expand the choice set of the model to include a larger set of metropolitan areas and gain more explanatory power by employing the vast LEHD panel dataset, which tracks the migration decisions and work histories of millions of individuals in the United States. In addition, I plan to perform counterfactual simulations using my estimated model to better understand the role of amenity values in both the location decision process and the observed wage differentials that exist across geographic locations.

Chapter 2

Using Panel Data to Easily Estimate Hedonic Demand Functions

2.1 Introduction

Property value hedonics uses the information in transaction prices of houses to recover the value of their non-marketed attributes. For decades, it has regularly been used as a tool to recover the value of amenities, including clean water and air, proximity to open space, and distance from toxic emitters and Superfund sites.[1] While this technique has become ubiquitous, it suffers from a number of well-known shortcomings. Foremost among these is a class of problems that arises when one attempts to use the hedonic approach to learn about individual preference heterogeneity and to measure the welfare consequences of non-marginal changes in amenities.

Rosen (1974) presents a two-step procedure for recovering the marginal willingness-to-pay (MWTP) function of a non-marketed amenity for an individual with a particular set of attributes. The first-stage of this procedure (*i.e.*, recovering the hedonic price gradient) is straightforward and can be used by itself to measure the value of marginal changes in amenities (or the value of non-marginal changes, as long as one makes strict assumptions about preference homogeneity). However, the econometrician typically faces a large bias due to omitted variables, as she is unable to observe all the features of a house and its neighborhood that are relevant to its price. Certain unobserved characteristics may be correlated with the environmental attributes that we seek to value. Short of simply collecting more data on housing characteristics, a solution to this problem has been found in using

[1]See Palmquist (2005) and Taylor (2003) for surveys of these applications, along with discussions of the property value hedonics methodology.

data on repeat housing sales, which allow all time-invariant attributes of a property to be controlled for non-parametrically with a series of fixed effects. See, for example, Palmquist (1982), Parsons (1992), and Mendelsohn et al. (1992).

The second-stage of Rosen's procedure uses hedonic prices (taken from the first-stage) as inputs in the estimation of functions describing demand for environmental attributes, which are then used to recover preference parameters for exact welfare measurement. The problems associated with this step are well documented. First among these is a simple problem of identification - without access to data from multiple markets, recovery of the MWTP function requires an untestable restriction on functional form [Brown and Rosen (1982), Mendelsohn (1985)]. More recently, Heckman, Ekeland, and Nesheim (2004) have shown that relaxing functional form restrictions in the recovery of the hedonic price function (*i.e.*, adopting a non-parametric estimation strategy) can avoid these problems.

Even after solving this fundamental identification problem, a second econometric complication remains in the latter half of Rosen's two-step procedure. When the hedonic price function is non-linear, home buyers simultaneously choose both the hedonic price and the quantity of the environmental amenity that they will consume. For example, when confronted with a convex, upward sloping hedonic price function (*e.g.*, with respect to air quality), a consumer with idiosyncratically strong preferences for clean air (*i.e.*, steep indifference curves) will choose to consume a lot of it at a high hedonic price, while a consumer with weak preferences (*i.e.*, flat indifference curves) will choose to consume a little at a low hedonic price [See Figure (2.1)]. The problem of consumer choice subject to a non-linear budget constraint creates a difficult endogeneity problem when using statistical inference to recover the parameters describing those preferences [Epple (1987), Bartik (1987)]. This issue has been addressed in previous research with a variety of instrumental variables strategies.[2] The problem with this solution is that it can be difficult to find valid, yet powerful,

[2]See, for example, Boyle, Poor, and Taylor (1999), Palmquist (1984), and Chattopadhyay (1999).

instruments. Typically, these IV approaches rely on some combination of the following assumptions: (i) certain socio-demographic variables enter directly into the MWTP function while others do not and the excluded variables can be used as instruments for endogenous attribute levels; (ii) the MWTP function only contains socio-demographic variables in linear form, while higher-order terms are excluded from the function and can therefore serve as instruments; or (iii) MWTP does not vary across markets, so that market dummies are excluded and can serve as instruments. In each case, these assumptions are not testable and place arbitrary restrictions on the estimated heterogeneity in MWTP.

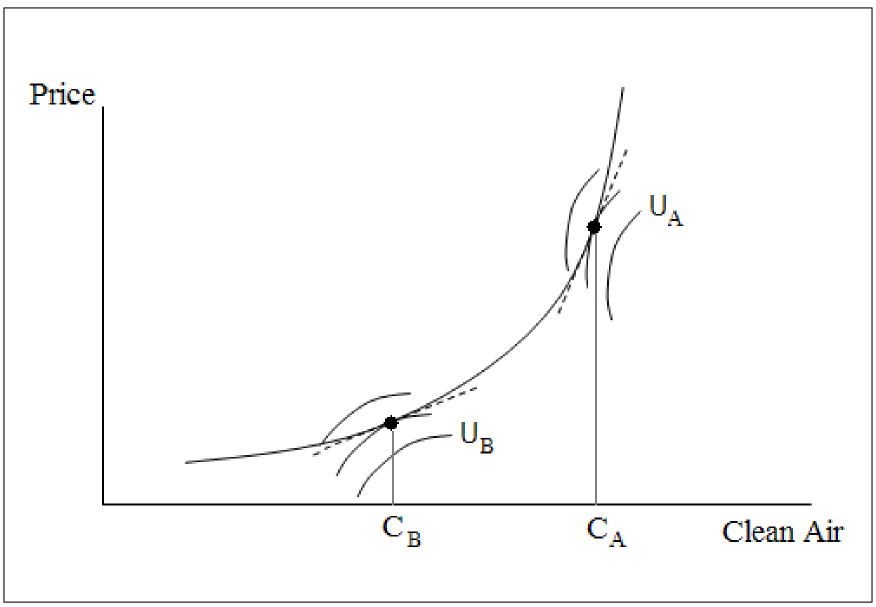

Figure 2.1: Preferences for Clean Air

Recent research has demonstrated that these problems can be avoided by replacing the statistical inference procedure in the second stage of Rosen's method with a "preference inversion" procedure, *i.e.*, specifying a functional form for individual utility that yields a set of first-order conditions for the optimal housing choice that can be solved for a unique set of preference parameters for each individual [Bajari and Benkard (2005)]. The strengths of this approach lie in (i) its admission of any form of preference heterogeneity across individuals and (ii) its avoidance of the endogeneity problems described above. Its primary weakness comes in the strict functional form assumptions that are required to perform the inversion procedure with typically available (*i.e.*, cross-sectional) data. In particular, one must assume that the MWTP function is flat, or restrict its elasticity to be -1.[3] When the shape of the MWTP function is so strongly dictated by functional form assumptions, the value of going beyond Rosen's first-stage becomes questionable. Bajari and Benkard recognize this, pointing-out that these assumptions can be relaxed if the researcher is able to observe the same individual home buyer on multiple purchase occasions. This is the starting point for this chapter.

In particular, we begin by using panel data to relax the strict assumptions required with cross-sectional data. We employ a "two-sided" housing panel, containing information on (i) repeat housing sales (*i.e.*, we see the same housing unit transact multiple times), and (ii) repeat housing purchases (*i.e.*, we observe the same individual purchase several housing units). We first show how repeat sales data can be used to overcome omitted variables bias, even in the context of the non-parametric first-stage estimation required by Bajari and Benkard. In particular, we employ non-parametric fixed effect techniques developed by Ullah and Roy (1998) in order to estimate a flexible set of hedonic price gradients while still controlling non-parametrically for time-invariant housing attributes. Previous research using repeat housing sales has relied on strict functional form assumptions to recover the hedonic price function. Second, we demonstrate that by observing each home buyer on at least two choice occasions, we can relax the assumptions made by Bajari and Benkard (using cross-sectional data) so as to be consistent with the functional form restrictions usually imposed under the Rosen two-step procedure.

We apply this methodology to data describing housing transactions and air quality in

[3]These results are illustrated in Section 2.2.2 and in the Appendix.

California's Bay Area. We find that a significant amount of heterogeneity exists, with support for preference-based sorting on ozone. Thus, the use of average MWTP would under-estimate the benefits of reducing ozone pollution in the worst areas.

2.2 Model

In this section, we describe our model. First, we discuss our non-parametric repeat sales model that we use to recover the hedonic gradient with respect to ozone pollution. Second, we describe Bajari and Benkard's approach to recovering heterogeneous preferences in a second stage preference inversion with cross-sectional data. Finally, we describe our second-stage approach that recovers inverse demand functions using a panel of buyers.

2.2.1 A Non-Parametric Fixed-Effects Approach to Recovering the Hedonic Gradient

The two-step hedonic approach described by Rosen (1974) begins with the estimation of a linear hedonic price function, which relates the price of house j transacted in period t ($P_{j,t}$) to its attributes, both those that vary over time ($Z_{j,t}$) and those that do not (X_j):[4]

$$P_{j,t} = X_j'\gamma + Z_{j,t}'\beta + \epsilon_{j,t} \tag{2.1}$$

Bias from omitted variables arises when $\text{E}[Z_{j,t} \cdot \epsilon_{j,t}] \neq 0$ or $\text{E}[X_j \cdot \epsilon_{j,t}] \neq 0$. This would be the case if data describing important neighborhood or housing attributes (*e.g.*, distance to city-center or curb-appeal) were not available to the researcher, but those variables were correlated with the attribute of interest (*e.g.*, air quality).

Panel data describing repeat sales of each house allows for time-invariant unobservables to be controlled for non-parametrically. To see this, decompose the error in Equation (2.1) into time-varying ($\nu_{j,t}$) and time-invariant (ξ_j) components:

[4]While the typical application of Rosen's procedure is to employ data from multiple markets, we illustrate it instead in a context with multiple time periods. This corresponds directly to our data and modeling strategy described below.

$$P_{j,t} = X_j'\gamma + Z_{j,t}'\beta + \xi_j + \nu_{j,t} \tag{2.2}$$

For each house j, differencing away the mean value of each variable (taken over all the sales of that house) controls for the influence of ξ_j on price:

$$\underbrace{P_{j,t} - \overline{P}_j}_{\tilde{P}_{j,t}} = \underbrace{(X_j - \overline{X}_j)'}_{0}\gamma + \underbrace{(Z_{j,t} - \overline{Z}_j)'}_{\tilde{Z}_{j,t}}\beta + \underbrace{(\xi_j - \overline{\xi}_j)}_{0} + \underbrace{(\nu_{j,t} - \overline{\nu}_j)}_{\tilde{\nu}_{j,t}} \tag{2.3}$$

where

$$\bar{P}_j = \frac{1}{T_j}\sum_t P_{j,t} \qquad \bar{Z}_j = \frac{1}{T_j}\sum_t Z_{j,t} \qquad \bar{\nu}_j = \frac{1}{T_j}\sum_t \nu_{j,t} \tag{2.4}$$

$$\bar{X}_j = \frac{1}{T_j}\sum_t X_j = X_j \qquad \bar{\xi}_j = \frac{1}{T_j}\sum_t \xi_j = \xi_j$$

T_j represents the number of times house j was sold. This yields the following equation, which can be taken to the data:

$$\tilde{P}_{j,t} = \tilde{Z}_{j,t}'\beta + \tilde{\nu}_{j,t} \tag{2.5}$$

Assuming $\mathrm{E}[\tilde{Z}_{j,t} \cdot \tilde{\nu}_{j,t}] = 0$, estimation of Equation (2.5) by ordinary least squares yields an unbiased estimate of β.

The simple linear framework described in Equations (2.1) - (2.5) imposes unrealistic restrictions on the equilibrium underlying the hedonic price function. Moreover, application of the techniques proposed by Bajari and Benkard (which we describe in Section 2.2.2) for the recovery of preference parameters generally requires a non-parametric representation of the hedonic price function. Fortunately, Ullah and Roy (1998) provide a non-parametric approach to the estimation of Equation (2.5) based on the idea of local linear regression.[5]

[5] See Fan and Gijbels (1996) for a description of local-linear regression.

Local linear regression relies on simple ordinary least squares estimation techniques. Rather than estimating a single set of OLS parameters, however, it recovers a different parameter vector for every point in the space of $Z_{j,t}$, weighting more heavily the information contained in data points that are "close" to $Z_{j,t}$.[6] As such, it has the ability to flexibly approximate almost any continuous function.

Following Ullah and Roy (1998), we begin by writing down a flexible representation of the hedonic price function.

$$P_{j,t} = f(Z_{j,t}) + \xi_j + \nu_{j,t} \tag{2.6}$$

where $f(\cdot)$ is an unspecified, flexible function of time-varying attributes of house j (or its neighborhood) and ξ_j represents all time invariant attributes (whether they are observed by the researcher or not). ν_j is assumed to be distributed *i.i.d.* with mean zero and constant variance σ_ν^2. In our application, $Z_{j,t}$ will consist of (i) a measure of ozone pollution at house j in year t (*e.g.*, the number of days in the course of each year that the state maximum 1-hour ozone concentration was violated) and (ii) the year of the housing transaction.[7]

Next, we take a first-order Taylor series expansion of $f(\cdot)$ around some vector χ (a vector with the same dimension as $Z_{j,t}$):[8]

$$P_{j,t} = f(\chi) + (Z_{j,t} - \chi)' f'(\chi) + \xi_j + \nu_{j,t} \tag{2.7}$$

We then consider the mean (taken over all sales of house j) of Equation (2.7):

[6] The definition of "close" is a decision made by the researcher, in her choice of kernel bandwidth. This is discussed in more detail below.

[7] Including the year that the house transacts controls for the rapid depreciations and appreciations observed over the period of our panel in the Bay Area. Other potential measures of ozone concentration include the 1-hour and 8-hour average maximum concentrations over the course of each year.

[8] The remainder term associated with the Taylor expansion is ignored.

$$\bar{P}_{j,t} = f(\chi) + (\bar{Z}_j - \chi)' f'(\chi) + \xi_j + \bar{\nu}_{j,t} \tag{2.8}$$

where

$$\bar{P}_j = \frac{1}{T_j}\sum_t P_{j,t} \qquad \bar{Z}_j = \frac{1}{T_j}\sum_t Z_{j,t} \qquad \bar{\xi}_j = \frac{1}{T_j}\sum_t \xi_{j,t} = \xi_j \qquad \bar{\nu}_j = \frac{1}{T_j}\sum_t \nu_{j,t} \tag{2.9}$$

Finally, taking within-house deviations from means allows us to difference away the fixed effect, ξ_j, and the function $f(\cdot)$.[9]

$$P_{j,t} - \bar{P}_j = (Z_{j,t} - \bar{Z}_j)' f'(\chi) + (\nu_{j,t} - \bar{\nu}_j) \tag{2.10}$$

Denoting differences-from-means with "$\sim$" and replacing $f'(\cdot)$ with $\beta(\cdot)$, we have:

$$\tilde{P}_{j,t} = \tilde{Z}'_{j,t}\beta(\chi) + \tilde{\nu}_{j,t} \tag{2.11}$$

Ullah and Roy (1998) show that $\beta(\chi)$ (*i.e.*, the slope of the hedonic price function at $Z_{j,t} = \chi$) can be recovered with the following minimization procedure:

$$\beta(\chi) = \underset{\beta(\chi)}{\operatorname{argmin}} \sum_{j=1}^{J}\sum_{t=1}^{T_j}\left(\tilde{P}_{j,t} - \tilde{Z}'_{j,t}\beta(\chi)\right)^2 K_h(Z_{j,t} - \chi) \tag{2.12}$$

where $\beta(\chi)$ is evaluated at values of χ equal to all of the observed values of $Z_{j,t}$. We therefore end up with a (potentially) different estimate of $\beta(\chi)$ at each data point; this is in contrast to the simple linear estimation in Equation (2.5), where β was constrained to be the same for every value of $Z_{j,t}$.

[9]Losing $f(\cdot)$ from this expression does not pose a problem, as our interest is only in recovering the hedonic gradient, *i.e.*, the slope of the hedonic price function, which is represented non-parametrically by $f'(\cdot)$.

Important to the local-linear regression procedure is the choice of weights placed on data as one moves further from χ. $K_h(\cdot)$ is the Gaussian kernel:

$$K_h(Z_{j,t} - \chi) = \prod_{k=1}^{\#regressors} \frac{1}{h\hat{\sigma}_{Z_k}} \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{1}{2}\left(\frac{Z_{k,j,t} - \chi_k}{h\hat{\sigma}_{Z_k}}\right)^2 \right\} \tag{2.13}$$

where h represents the kernel bandwidth and $\hat{\sigma}_{Z_k}$ is the standard deviation of the k^{th} element of $Z_{j,t}$. After visual inspection of the hedonic gradient, we choose this bandwidth to be $5.3 \cdot n^{-\frac{1}{6}}$, where n is the number of observations.[10] Given the linear nature of the problem, there exists a closed-form solution for $\beta(\chi)$:

$$\beta(\chi) = (\tilde{Z}'W_\chi \tilde{Z})^{-1}\tilde{Z}'W_\chi \tilde{P} \tag{2.14}$$

where $\tilde{Z}$ is an (n x k) matrix of mean-differenced regressors, $\tilde{P}$ is an (n x 1) vector of mean-differenced house prices, and W_χ is an (n x n) matrix of weights, such that $W_\chi = diag(K_h(Z_{j,t} - \chi))$.

Since we allow χ to take on each observed value of the vector Z, the output of this local-linear regression procedure provides us with an estimate of $\beta(\chi)$ (*i.e.*, the fully flexible slope of the hedonic price function) at each observed value of Z. We next take this rich information on hedonic prices to the identification of individual preference parameters.

2.2.2 Recovering Heterogeneous Preferences with Cross-Sectional Data

In contrast to the canonical Rosen two-step approach (in which hedonic prices are recovered from the estimated gradient and used in a second-stage estimation to recover MWTP functions), the approach presented in Bajari and Benkard begins by specifying the

[10] Our bandwidth is set to five times the value suggested by Silverman's Rule of Thumb, which suggests setting the bandwidth value to $1.06 \cdot n^{-\frac{1}{6}}$. We find that, in practice, this tends to under-smooth the estimated hedonic price function. [Silverman (1986)]

functional form for the utility function.[11] Suppose individual i in period t chooses a house to maximize indirect utility over a time-varying amenity such as air quality ($z_{j,t}$) and other (fixed) house attributes (X_j). In the case of cross-sectional data (in which each home buyer i is observed only once and z is not subscripted by t), one possible specification for utility is:[12]

$$U(X_j, z_j, C_i) = \phi_{X,i} X_j + \phi_{z,i} z_j + \phi_{C,i} C_i \tag{2.15}$$

Where X_j and z_j represent attributes of house j and C_i represents i's consumption of other non-housing goods. We use a monotonic transformation of utility to normalize $\phi_{C,i} = 1$. Incorporating individual i's budget constraint, $R(X_j, z_j) + C_i = I_i$, yields the indirect utility function:

$$V(X_j, z_j, I_i) = \phi_{X,i} X_j + \phi_{z,i} z_j + [I_i - R(X_j, z_j)] \tag{2.16}$$

where $R_{j,t}$ is the imputed annual rent or housing expenditure. In practice, we calculate this figure as 7.5% of the observed transaction price.[13] As $R_{j,t}$ is a constant fraction of $P_{j,t}$, the following relationship holds: $\frac{\partial R_{j,t}}{\partial z_{j,t}} = (0.075) \cdot \frac{\partial P_{j,t}}{\partial z_{j,t}} = \rho_{j,t}$.

The first-order condition corresponding to the optimal choice of z_j is given by:

$$\phi_{z,i} - \frac{\partial R(X_j, z_j)}{\partial z_j} = 0 \tag{2.17}$$

Individual i's marginal utility of z is revealed by the hedonic price at the value $z_{j^*(i)}$

[11] The Rosen two-step procedure instead specifies a functional form for the MWTP function.

[12] Bajari and Benkard explicitly include the unobserved attribute, ξ_j, among the arguments of the utility function. For our purposes, it is sufficient to think of the unobserved housing attribute ξ_j as being included in the X_j vector.

[13] This is the commonly used discount in the literature and is estimated in Peiser and Smith (1985).

actually chosen by the individual:[14]

$$\phi_{z,i} = \rho_{j^*(i)} = \frac{\partial R(X_j, z_j)}{\partial z_j}\bigg|_{z_j = z_{j^*(i)}} \tag{2.18}$$

The restriction imposed on utility in order to be able to recover preferences in the preceding example is that the marginal utility of z_j is constant, implying that the MWTP for z_j is a horizontal line that does not depend upon income or on the quantity of z_j consumed. Equation (2.18) then constitutes a single equation in a single unknown parameter, $\phi_{z,i}$.

While it is a desirable feature of this model that MWTP can vary in an unrestricted fashion across individuals (the same is not true when we simply recover the average MWTP from the hedonic price function), it is unrealistic to expect that MWTP would not vary with the quantity of z_j consumed.[15] This has the potential to severely bias estimates of the welfare effects of non-marginal changes in z_j (*i.e.*, leading to an understatement of the benefits associated with a reduction in (the undesirable) attribute z_j).

2.2.3 Recovering a Flexible MWTP Function with a Panel of Buyers

In this section, we demonstrate that it is straightforward to recover the flexible distribution of linear MWTP functions by employing the techniques utilized by Bajari and Benkard in conjunction with panel data on home purchasers. We begin by specifying the indirect utility of individual i choosing a home j with attributes $(X_j, z_{j,t})$ in period t:

$$V_{i,j,t} = \alpha_{0,i} + \alpha_{1,i}X_j + \alpha_{2,i}X_j^2 + \alpha_{3,i}z_{j,t} + \alpha_{4,i}z_{j,t}^2 + (I_{i,t} - R_{j,t}) \tag{2.19}$$

We now demonstrate that this functional form will yield the same linear MWTP specifi-

[14]See Bajari and Kahn (2005) for a practical approach when $z_{j^*(i)}$ is restricted to discrete values and the first order condition in Equation (2.17) need not hold exactly.

[15]Using only functional form restrictions, one can recover a demand curve using techniques in Bajari and Benkard that varies with z_j, but in a highly restricted fashion. See the Appendix for a derivation.

cation that is common in the hedonics literature. Moreover, as long as individual attributes are not time-varying, the parameters of the demand function can be identified with just two observations for each home buyer. Consider the first-order conditions associated with choosing the optimal value of $z_{j,t}$:

$$\frac{\partial V_{i,j,t}}{\partial z_{j,t}} = \alpha_{3,i} + 2\alpha_{4,i} z_{j,t} - \frac{\partial R_{j,t}}{\partial z_{j,t}} = 0 \tag{2.20}$$

For individual i, we need to recover the values of two unknown parameters, $(\alpha_{3,i}, \alpha_{4,i})$.[16] Fortunately, we observe individual i in panel data on (at least) two occasions. This gives us two equations in two unknowns:

$$\alpha_{3,i} + 2\alpha_{4,i} z_{j^*(i),1} = \rho_{j^*(i),1} \qquad \alpha_{3,i} + 2\alpha_{4,i} z_{j^*(i),2} = \rho_{j^*(i),2} \tag{2.21}$$

where $\rho_{j^*(i),t} = \frac{\partial R_{j,t}}{\partial z_{j,t}}\Big|_{z_{j,t}=z_{j^*(i),t}}$.

Solving these two equations yields:

$$\hat{\alpha}_{3,i} = \frac{\rho_{j^*(i),1} z_{j^*(i),2} - z_{j^*(i),1} \rho_{j^*(i),2}}{z_{j^*(i),2} - z_{j^*(i),1}} \qquad \hat{\alpha}_{4,i} = \frac{1}{2} \frac{\rho_{j^*(i),2} - \rho_{j^*(i),1}}{z_{j^*(i),2} - z_{j^*(i),1}} \tag{2.22}$$

Having recovered $(\widehat{\alpha}_{3,i}, \widehat{\alpha}_{4,i})$, we can then recover individual i's MWTP function for z:

$$\rho_{j^*(i),t} = \hat{\alpha}_{3,i} + 2\hat{\alpha}_{4,i} z_{j^*(i),t} \tag{2.23}$$

[16]Murray (1975) takes the approach of using an equation very similar to (3.20), except without individual heterogeneity in preference parameters, to form an estimating equation. Recovering estimates of $\frac{\partial R_{j,t}}{\partial z_{j,t}}$ from the first-stage, he then recovers estimates of $\widehat{\alpha}_3, \widehat{\alpha}_4$ from a two-stage least squares procedure, using income and prices as instruments.

Finally, we can determine how these inverse demand functions differ systematically with fixed individual attributes (A_i) by performing the following regressions:

$$\hat{\alpha}_{3,i} = \delta_{3,0} + A_i'\delta_{3,1} + \varsigma_{3,i} \qquad \hat{\alpha}_{4,i} = \delta_{4,0} + A_i'\delta_{4,1} + \varsigma_{4,i} \tag{2.24}$$

Adding Time-varying Individual Attributes

An implication of the preceding specification is that MWTP is not allowed to vary with time-varying observable purchaser attributes, *e.g.*, expenditure on non-housing consumption. Income is often included as a determinant of MWTP, but is time-varying (in panel data) and hence cannot be summarized by the vector of fixed attributes, A_i. Murray (1983) shows why it is important for the MWTP function to vary with non-housing consumption expenditures. We demonstrate next how time varying non-housing consumption expenditures can be included, as long as the researcher has *three* observations on each individual. We begin by specifying individual i's indirect utility from choosing house j in period t as:

$$V_{i,j,t} = \alpha_{0,i} + \alpha_{1,i}X_j + \alpha_{2,i}X_j^2 + \alpha_{3,i}z_{j,t} + \alpha_{4,i}z_{j,t}^2 + \alpha_{5,i}z_{j,t}(I_{i,t} - R_{j,t}) \tag{2.25}$$

Now, we allow the marginal utility of $z_{j,t}$ to vary with non-housing consumption expenditure. The first-order condition associated with the individual's optimal choice of $z_{j,t}$ is then given by:

$$\frac{\partial V_{i,j,t}}{\partial z_{j,t}} = \alpha_{3,i} + 2\alpha_{4,i}z_{j,t} + \alpha_{5,i}(I_{i,t} - R_{j,t}) \underbrace{-\alpha_{5,i}z_{j,t}\frac{\partial R_{j,t}}{\partial z_{j,t}} - \frac{\partial R_{j,t}}{\partial z_{j,t}}}_{-\rho_{j,t}(1+\alpha_{5,i}z_{j,t})} = 0 \tag{2.26}$$

Presuming that we observe individual i on three separate purchase occasions, Equation (2.26) yields three equations in as many unknowns ($\alpha_{3,i}, \alpha_{4,i}, \alpha_{5,i}$):

$$\alpha_{3,i} + 2\alpha_{4,i}z_{j^*(i),1} + \alpha_{5,i}(I_{i,1} - R_{j^*(i),1}) - \rho_{j^*(i),1}(1 + \alpha_{5,i}z_{j^*(i),1}) = 0 \tag{2.27}$$

$$\alpha_{3,i} + 2\alpha_{4,i}z_{j^*(i),2} + \alpha_{5,i}(I_{i,2} - R_{j^*(i),2}) - \rho_{j^*(i),2}(1 + \alpha_{5,i}z_{j^*(i),2}) = 0$$

$$\alpha_{3,i} + 2\alpha_{4,i}z_{j^*(i),3} + \alpha_{5,i}(I_{i,3} - R_{j^*(i),3}) - \rho_{j^*(i),3}(1 + \alpha_{5,i}z_{j^*(i),3}) = 0$$

Solving this system provides us with the following equations for individual i's preference parameters:

$$\hat{\alpha}_{4,i} = \frac{\theta_{j^*(i),3}(\rho_{j^*(i),1} - \rho_{j^*(i),2}) - \theta_{j^*(i),2}(\rho_{j^*(i),1} - \rho_{j^*(i),3})}{2\theta_{j^*(i),3}(z_{j^*(i),1} - z_{j^*(i),2}) - 2\theta_{j^*(i),2}(z_{j^*(i),1} - z_{j^*(i),3})} \tag{2.28}$$

$$\hat{\alpha}_{5,i} = \frac{2\hat{\alpha}_{4,i}(z_{j^*(i),1} - z_{j^*(i),3}) - \rho_{j^*(i),1} + \rho_{j^*(i),3}}{\rho_{j^*(i),1}z_{j^*(i),1} - \rho_{j^*(i),3}z_{j^*(i),3} + (I_{i,3} - R_{j^*(i),3}) - (I_{i,1} - R_{j^*(i),1})}$$

$$\hat{\alpha}_{3,i} = \rho_{j^*(i),1}(1 + \alpha_{5,i}z_{j^*(i),1}) - 2\hat{\alpha}_{4,i}z_{j^*(i),1} - \hat{\alpha}_{5,i}(I_{i,1} - R_{j^*(i),1})$$

where

$$\theta_{j^*(i),2} = \rho_{j^*(i),1}z_{j^*(i),1} - \rho_{j^*(i),2}z_{j^*(i),2} + (I_{i,2} - R_{j^*(i),2}) - (I_{i,1} - R_{j^*(i),1}) \tag{2.29}$$

$$\theta_{j^*(i),3} = \rho_{j^*(i),1}z_{j^*(i),1} - \rho_{j^*(i),3}z_{j^*(i),3} + (I_{i,3} - R_{j^*(i),3}) - (I_{i,1} - R_{j^*(i),1})$$

Having solved for all three parameters, individual i's MWTP function for Z is given by:

$$\rho_{j^*(i),t} = \frac{\hat{\alpha}_{3,i} + 2\hat{\alpha}_{4,i}z_{j^*(i),t} + \hat{\alpha}_{5,i}(I_{i,t} - R_{j^*(i),t})}{(1 + \hat{\alpha}_{5,i}z_{j^*(i),t})} \tag{2.30}$$

Summary

Given a sufficient number of repeat purchase observations, the techniques outlined above allow for the recovery of a rich distribution of MWTP functions with shape restrictions that are typical of those used in most Rosen two-step applications. Following Bajari and Benkard, we obtain these by inverting systems of first-order conditions for each individual home-buyer and introducing information from the non-parametric first-stage estimation of the hedonic gradient. The use of this panel data inversion procedure allows the difficult endogeneity issues associated with the Epple (1987) critique to be avoided, while also avoiding strong restrictions on preferences. Moreover, preferences can be allowed to vary over time with observable individual attributes (*e.g.*, non-housing consumption expenditure, number of children, employment location), by simply using additional repeat purchases.

2.3 Data

Implementing our panel data variant of the Bajari and Benkard estimator requires data describing multiple housing purchase decisions by a single individual. Adequately controlling for house attributes in the first stage hedonic regression, moreover, requires data on multiple sales per house. We assemble such a dataset by combining information from a real estate transactions dataset and a dataset describing mortgage applicants' demographic characteristics obtained through the Home Mortgage Disclosure Act (HMDA).

2.3.1 Property Transactions Data

The real estate transactions data we employ cover the six core counties of the San Francisco Bay Area (Alameda, Contra Costa, Marin, San Francisco, San Mateo, and Santa Clara) over the period 1990 to 2004. The data are purchased from Dataquick and include transaction dates, prices, loan amounts, and buyers', sellers' and lenders' names for all transactions. In addition, the data for the final observed transaction include housing characteristics, such as exact street address, square footage, year built, lot size, number of rooms, number of bathrooms, and number of units in the building.

As our analysis differences out all non time varying housing characteristics (and housing attributes are available in the data only for the most recent transaction), it is necessary to ensure we are comparing the same house over time. First, to control for land sales or re-builds, we drop all transactions where "year built" is missing or with a transaction date that is prior to "year built." Second, we want to control for property improvements (*e.g.*, an updated kitchen) or degradations (*e.g.*, water damage) that would not present as a re-build. In a repeat-sales analysis, similar in spirit to Case and Shiller (1989), we regress log prices on a set of house and year dummies (omitting the dummy for the base year, 1990). This gives us a crude measure of yearly appreciation rates in the Bay Area. In years where there is an overall appreciation in housing values (1995-2004), we drop properties that experience a yearly appreciation that is more than five times the average appreciation or properties that experience an average depreciation that is sixty percentage points lower than the average appreciation. In years where there was an overall depreciation in housing values (1990-1994), we drop properties that experience a yearly depreciation that is more than four times higher (in absolute value) than average depreciation or properties that experience an appreciation more than sixty percentage points higher than the average depreciation. In addition, we look at movement within the overall distribution of prices, dropping properties that move more than thirty percentage points in either direction between observed transactions.

Additionally, we drop transactions where the price is missing, negative, or zero. After using the consumer price index to convert all transaction prices into 2000 dollars, we drop one percent of observations from each tail to minimize the effect of outliers. Finally, as we merge in the pollution data using the property's geographic coordinates, we drop properties where latitude and longitude are missing.

Finally, we restrict the sample to properties that sold at least twice. We also drop properties that sell more than once within a given year or more than five times over the fifteen year period. This yields a final sample of 438,492 transactions (*i.e.*, property-year observations), comprised of 191,210 unique properties. Table (2.1) compares the set of repeat sales houses to the full sample of homes in the data. In most dimensions, the samples are very similar.

Table 2.1: Repeat Sale Sample vs. Full Sample

	Full Sample *properties = 594,665*		Repeat-Sale Sample *properties = 191,210*	
variable	mean	median	mean	median
Price (2000 $)	377,684	330,000	377,158	336,673
Sq. Ft. House	1,696.95	1,500.00	1,589.71	1,432.00
Sq. Ft. Lot	9,611.62	5,500.00	8,384.23	5,044.00
Year Built	1966.48	1970	1968.60	1972
Total Rooms	6.38	6	6.21	6
Num. Bedrooms	2.94	3	2.86	3
Num. Bath	2.02	2	1.99	2
Days $\geq$ 0.09 ppm Ozone	2.44	2.09	2.54	2.09

2.3.2 Buyer Characteristics Data

In order to implement the panel data version of the Bajari and Benkard estimator, we need to create a panel of buyers, their characteristics, and their chosen properties over time. This involves first identifying and matching individuals over time in the property transactions dataset and, second, merging in the individual attribute data from the HMDA dataset. We use the algorithm found in Bayer, McMillan, Murphy, and Timmins (2008).

To link individuals over time within the transactions data, the algorithm matches on the individual's name and the date of the transaction. See Figure (2.2). We first assign a unique identifier to all sellers. For instance, we assign the identifiers "A", "B", and "C" to observed sellers. We then look to that property's prior sale using the unique property id (and checking to see that individual's name is a match) and assign "A", "B", or "C" to the respective buyer. Next, given that we see individual "A" selling a property in period two, we look to see if "A" repurchases another property in period two.[17] If this is the case, and

[17]We look within a 24 month period, 12 months on either side of the sales date.

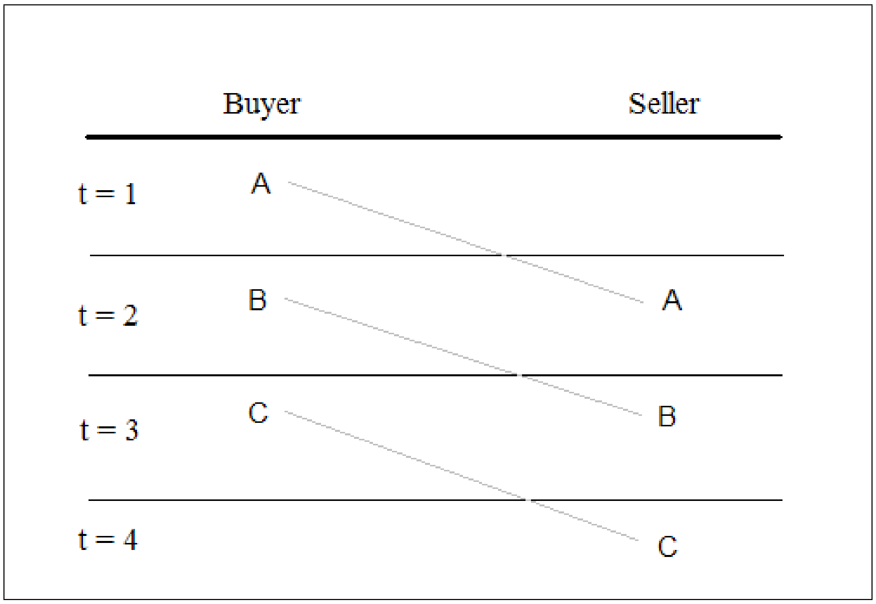

Figure 2.2: Tracking Buyers through Time

if buyer "B" and seller "A" have the same name, we would re-assign "B" as "A".[18]

The individual attribute data come from a dataset on mortgage applications published through the Home Mortgage Disclosure Act. These data provide information on all mortgage applications filed in the Bay Area over the relevant time period. Included are all applicants' race and gender, income, loan amount, lender name, and Census tract of the property.[19]

We are able to merge the individual attribute data in the HMDA dataset to the buyers and properties in the transactions dataset using the common variables of lender name, loan amount, transaction date, and Census tract of the property. We successfully match approximately 75% of individuals in the transactions sample to HMDA sample. Additionally, we drop individuals who purchase only once or more than three times over the sample period.

[18]We require that the last name matches exactly and the first name matches on the first letter.

[19]Census tracts are small, relatively homogenous geographic units defined by the Census Bureau. They contain between 2,500 and 8,000 individuals on average and vary in geographic size according to population density. [U.S. Census Bureau]

Finally, we drop individuals with missing income or race variables and observations with missing purchase prices. Again, as we merge pollution data using geographic coordinates, we drop observations where latitude and longitude are missing. This leaves us with a final sample of 9,805 individuals who purchase twice and 1,076 individuals who purchase three times. Table (2.2) compares the two samples of repeat purchasers with the full sample. As before, in most cases the samples are remarkably similar.

Table 2.2: Repeat Buyer Sample vs. Full Sample

	Full Sample *ind = 40,092*		Two-Purchase *ind = 9,805*		Three-Purchase *ind = 1,076*	
variable	mean	median	mean	median	mean	median
Asian	0.21	0	0.22	0	0.21	0
Black	0.02	0	0.02	0	0.02	0
Hispanic	0.09	0	0.10	0	0.10	0
White	0.59	1	0.58	1	0.58	1
Oth. Race	0.09	0	0.08	0	0.09	0
Income	123,033	98,993	124,805	101,324	131,637	101,826
Price	409,943	344,778	431,786	365,997	437,937	354,653

2.3.3 Ozone Data

The ozone data we employ are taken from the California Air Resources Board.[20] We use yearly ozone data from all thirty-seven monitors in the nine counties of Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma over the period 1990 to 2004. In particular, we use the monitor data to construct property-specific measures of the number of days exceeding the one-hour state standard (*i.e.*, 0.09 parts per million).

In addition to the ozone readings, the dataset provides information on the "year coverage," or the percent of time (during the relevant high-ozone season) each particular monitor was available and the geographic coordinates of each monitor. Using the coverage variable,

[20] Publicly available at www.arb.ca.gov/adam/.

we drop monitors with less than 60 percent coverage in a given year (amounting to less than 4 percent of the available monitor-year observations). Using the latitudinal and longitudinal coordinates of the monitors and the properties, we use the "Great Circle" estimator to compute the distance to all monitors from each property. We then create a weighted average for each property of all monitors' readings, using one over distance-squared as the weight. In order to mitigate the boundary effects, we include monitor data from the surrounding counties of Napa, Solano, and Sonoma, in addition to the six counties that appear in our transactions data. Figure (2.3) describes the ozone exceedance data.

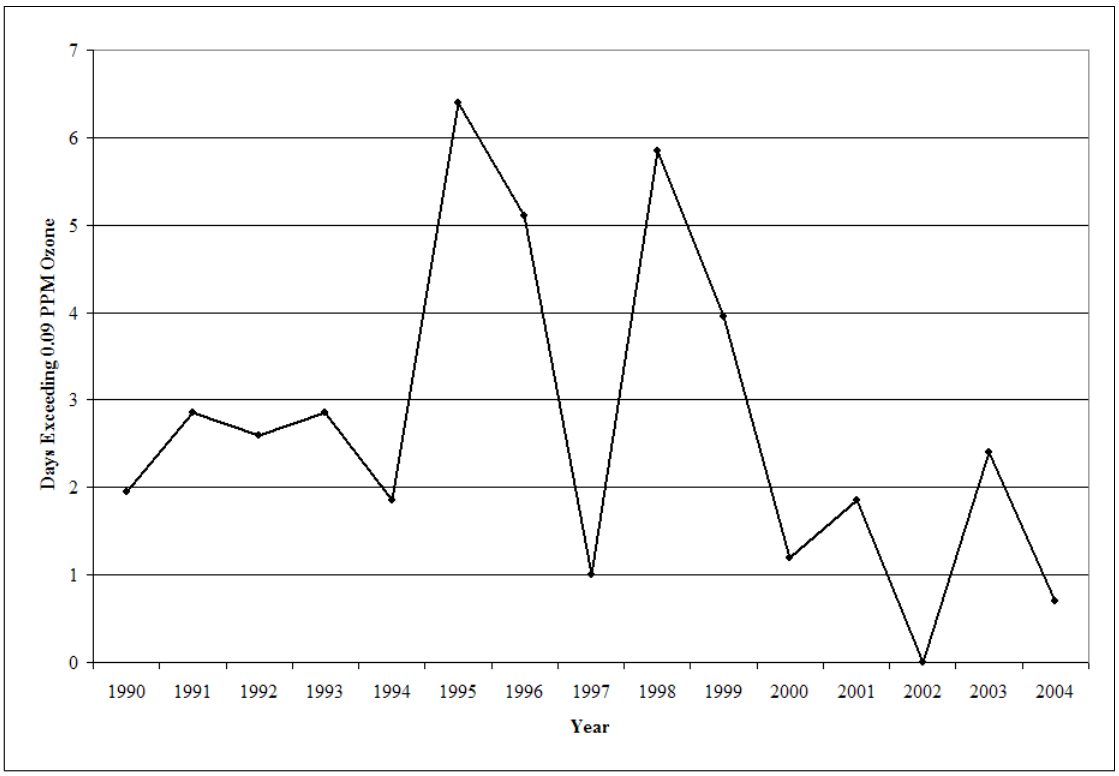

Figure 2.3: Average Number of Days Exceeding CA Ozone Limit

2.4 Results

In this section, we describe the results of our estimation procedure. First, we illustrate the results of our non-parametric repeat sales model that is used to recover the hedonic gradient with respect to ozone pollution. Second, we use the panel of 9,805 home buyers who are observed purchasing two houses in our data set to recover estimates of the inverse demand functions described above. Third, we use the sample of 1,076 individuals who buy three times to recover estimates of the inverse demand functions that are allowed to vary with non-housing expenditure. We also consider the average MWTP and the heterogeneous MWTP recovered using cross-sectional data.

2.4.1 Results from the First-stage Hedonic Regressions

The local linear estimation allows the estimate of the slope coefficient, $\beta(\chi)$, to differ for each observed value of ozone pollution in each year of the panel. Thus, for each year, the gradient can be graphically represented as a flexible function of ozone pollution. For ease of exposition, we show the hedonic gradient for three years of the panel (1992, 1997, and 2002) in Figure (2.4).

2.4.2 Results from the Second Stage Preference Inversion with Cross-Sectional Data

Using the full sample of observed buyers (treated as a cross-section), we perform the Bajari and Benkard preference inversion using the estimates from the first stage hedonic regressions. The trimmed distribution of MWTP is shown in Figure (2.5).[21]

The main benefit of the Bajari and Benkard approach is that the preference inversion generates a distribution of MWTP (although this comes at the cost of strict functional form assumptions). As can be seen in Figure (2.5), there is substantial heterogeneity in

[21] To limit the effect of outliers, 2.5% are dropped from either tail of the distribution.

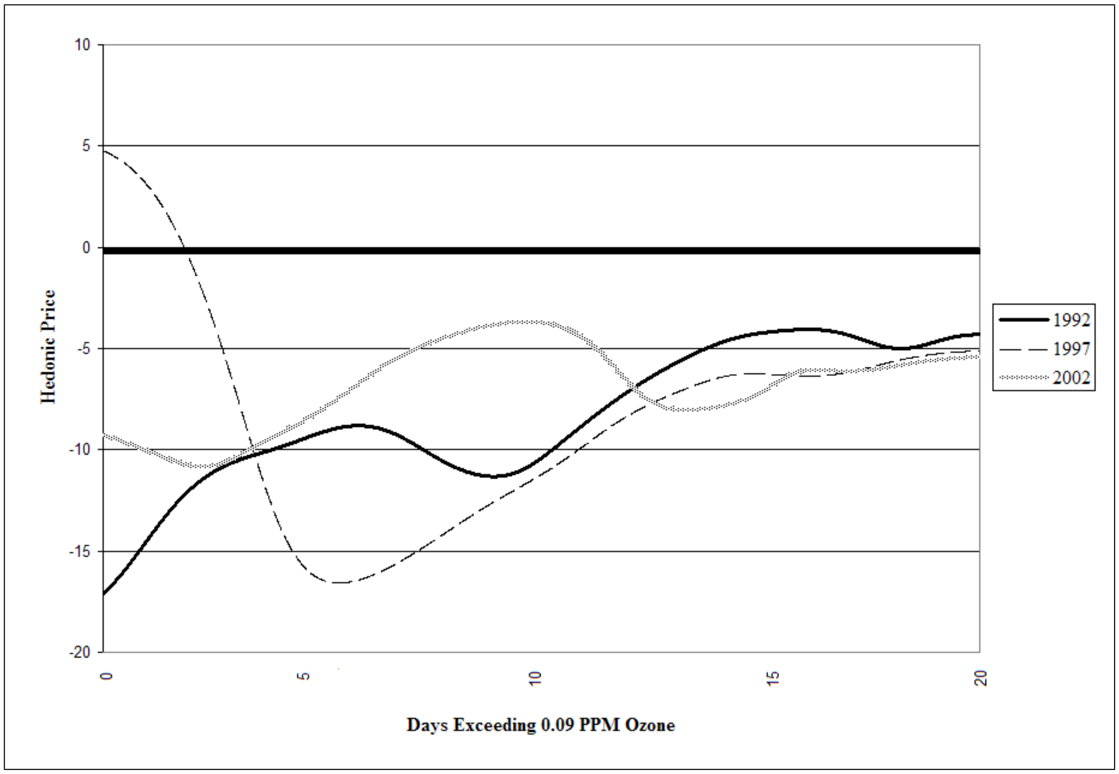

Figure 2.4: Hedonic Gradients from the Local Linear Estimation

MWTP. The distribution has a mean of −619.47 (or an average MWTP of $619.47 to avoid an additional day over the exceedance of ozone), a median of −721.50, and a standard deviation of 504.00.

Finally, we look for evidence of preference-based sorting by examining the relationship between MWTP and ozone pollution. In this analysis, the correlation between MWTP to avoid ozone pollution and the days exceeded variable is 0.25. This implies that individuals who live in high ozone areas are willing to pay more to avoid additional ozone pollution than individuals who currently live in low ozone areas. This means that the benefits associated with an ozone reduction in the dirtiest areas would be understated using average MWTP.

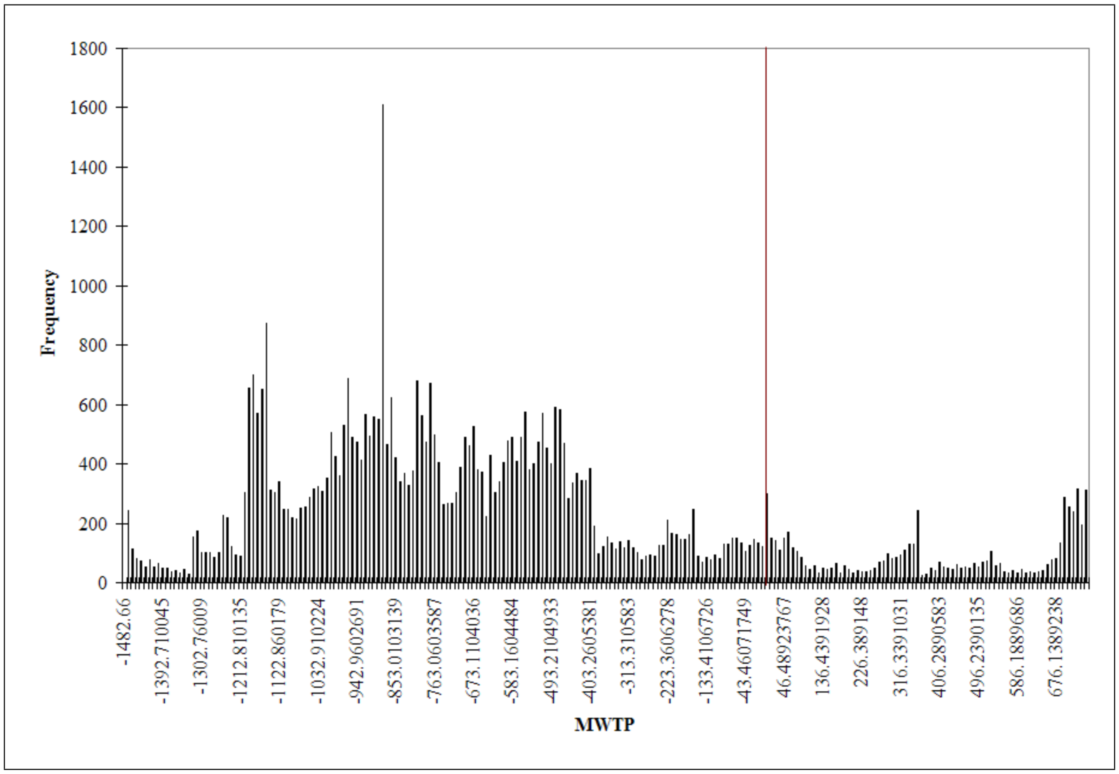

Figure 2.5: Heterogenous WTP using Bajari and Benkard Cross-Sectional Approach

Demand Estimation with Two-Purchase Panel

As previously discussed, employing the Bajari and Benkard preference inversion with a panel of buyers who are observed on two separate purchase occasions allows for the estimation of a linear MWTP function for each individual. Thus, MWTP is allowed to vary with the level of ozone pollution and we recover both an intercept and a slope coefficient for each individual. The trimmed distributions of these intercepts and slopes are shown in Figures (2.6) and (2.7), respectively.[22]

The distribution of intercepts shows significant heterogeneity. The distribution has a mean of -488.28, a median of -600.74, and a standard deviation of 1,040.81. While it is clear

[22]To limit the effect of outliers, 2.5% are dropped from either tail of the distribution.

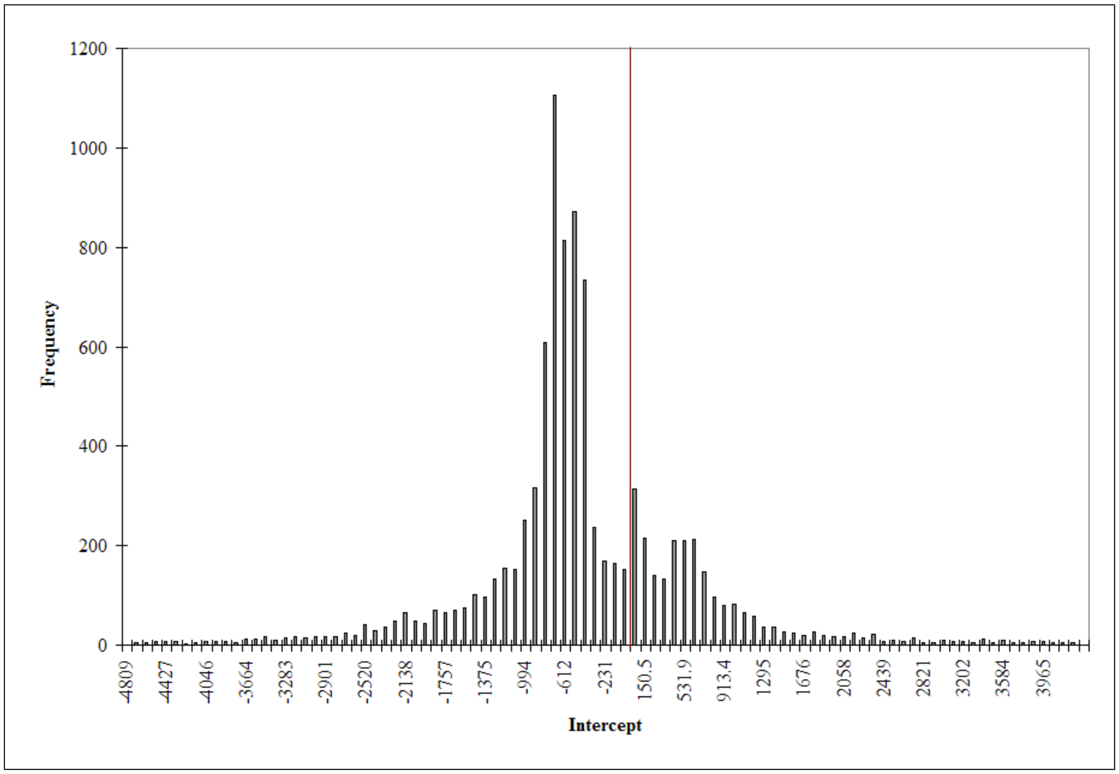

Figure 2.6: Distribution of Intercepts using Two-Purchase Panel

that the majority of intercepts are negative, the distribution of slope coefficients is centered closer to zero, with a mean of -1.09 and median of -26.92. The standard deviation of this distribution is 515.87, again implying that a substantial amount of heterogeneity exists. The correlation between the slope coefficient and level of ozone is 0.05, again implying that the individuals living in the higher ozone areas are more sensitive to changes in ozone (*i.e.*, their MWTP function is steeper). Thus, using a constant MWTP to calculate benefits could lead to a substantial downward bias, as it would mask considerable heterogeneity of preferences. Figure (2.8) shows estimated linear demand functions for three separate individuals.

Finally, we are able to calculate the inverse elasticity of ozone demand with respect to price. The trimmed distribution of these elasticities is shown in Figure (2.9). Again we see

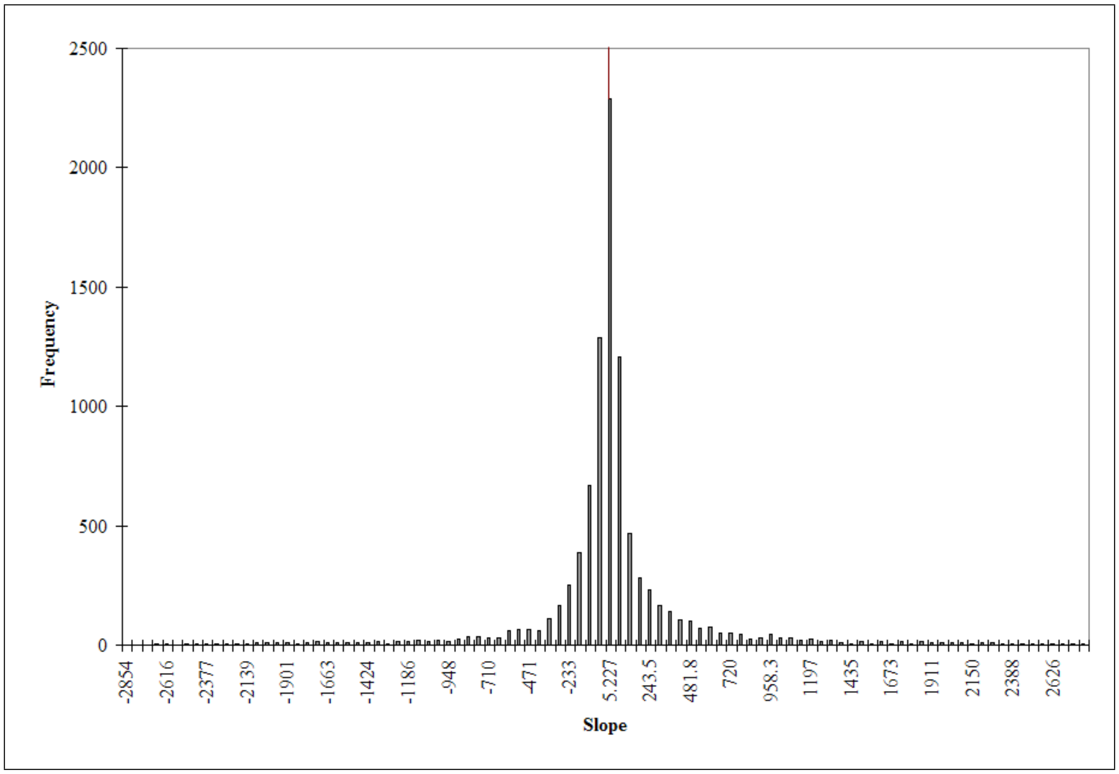

Figure 2.7: Distribution of Slope Coefficients using Two-Purchase Panel

considerable heterogeneity. As expected with inverse price elasticities, the distribution is centered close to zero, with a mean of 0.10, a median of 0.04, and a standard deviation of 1.82. In this case a positive elasticity means that the demand curves slope down (*i.e.*, the inverse elasticity $= \frac{\partial \rho_{j,t}}{\partial z_{j,t}} \cdot \frac{z_{j,t}}{\rho_{j,t}}$, where both the derivative and $\rho_{j,t}$ are negative).

Demand Estimation with Three-Purchase Panel

By using the sample of buyers who are observed to purchase on three separate occasions, we are able to allow MWTP to vary with income. For each individual, we are then able to calculate an income elasticity of demand. The trimmed distribution of these elasticities is plotted in Figure (2.10).

Although most of the elasticities are negative, implying that the demand curve shifts

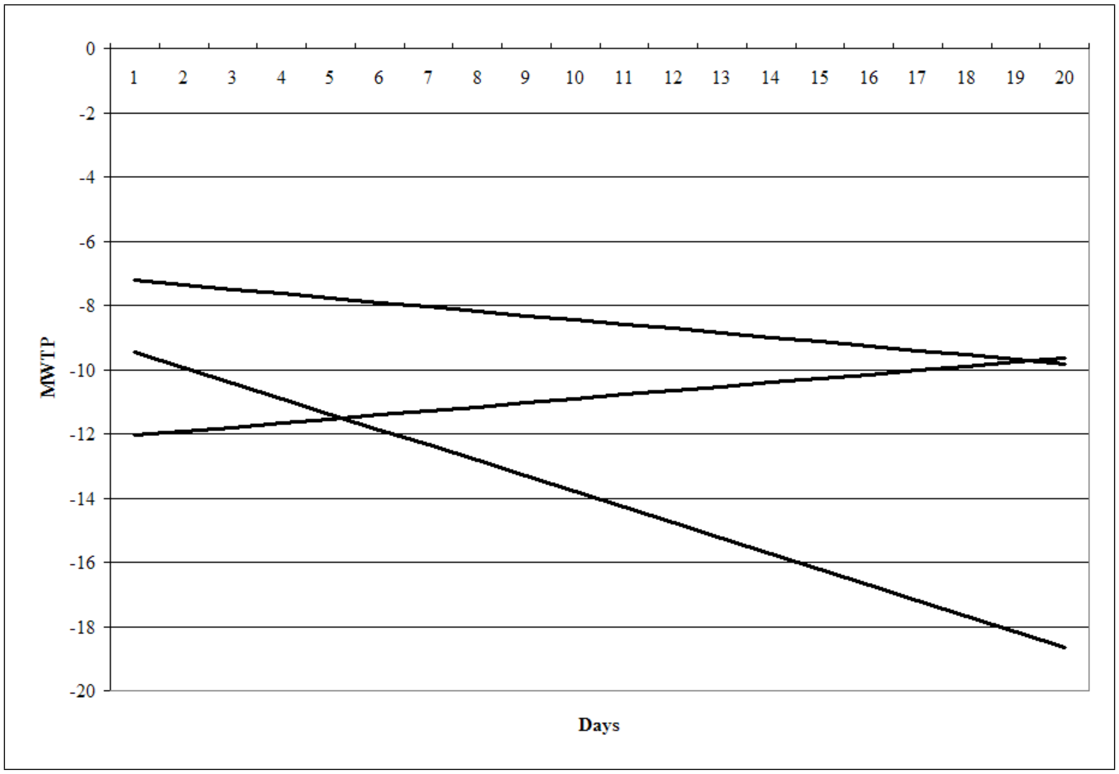

Figure 2.8: Individual Linear MWTP Functions

down when income goes up (and leads to a higher MWTP for a reduction in ozone), there is considerable heterogeneity and a wide distribution that includes many positive elasticities. The mean of the distribution is -1.43, with a median of -0.10 and a standard deviation of 33.53.

Finally, Figure (2.11) shows the trimmed distribution of inverse price elasticities. We again see considerable heterogeneity, with a mean of 0.15, a median of 0.04, and a standard deviation of 1.13. Again, in this case a positive elasticity means that the demand curves slope down (*i.e.*, the inverse elasticity $= \frac{\partial \rho_{j,t}}{\partial z_{j,t}} \cdot \frac{z_{j,t}}{\rho_{j,t}}$, where both the derivative and $\rho_{j,t}$ are negative).

We again see evidence of preference-based sorting in the correlation of 0.01 between the slope of the MWTP and level of pollution, with the MWTP function becoming steeper as

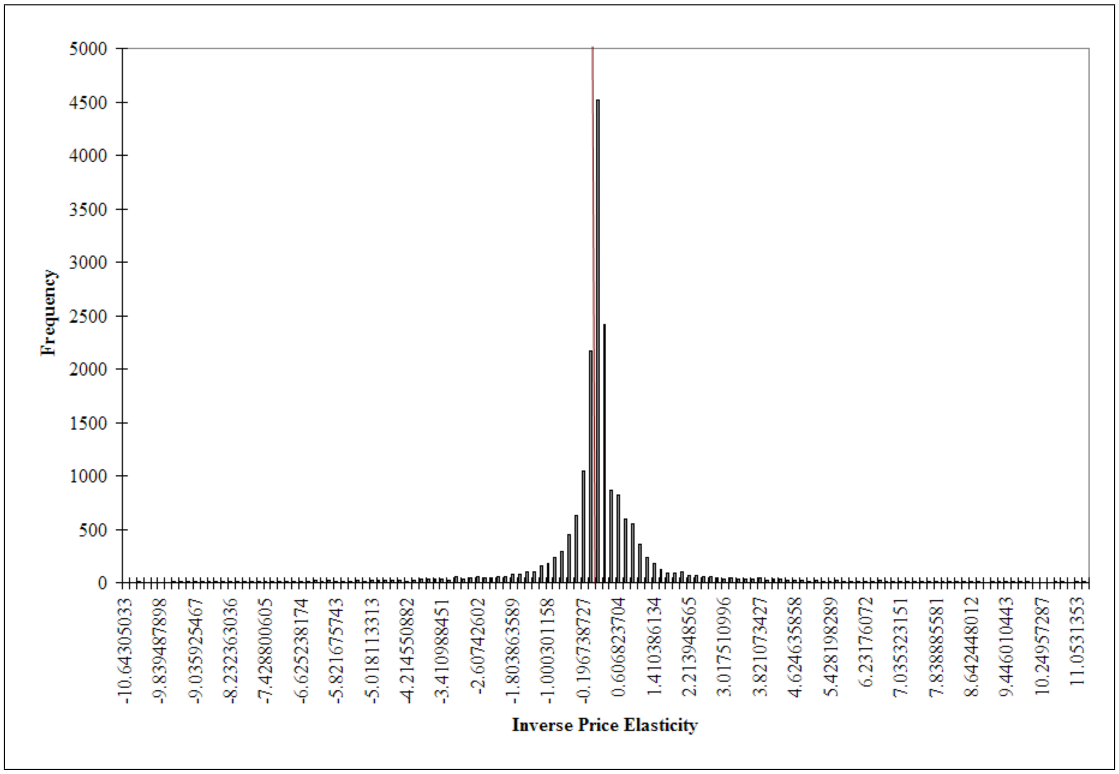

Figure 2.9: Distribution of Price Elasticities using Two-Purchase Panel

number of days above the ozone exceedance increases.

2.5 Conclusion

In order to measure the welfare consequences of non-marginal changes in amenity values, it becomes necessary to consider individual preference heterogeneity and the possibility of preference-based sorting. However, the standard framework for identifying heterogeneity in preferences, the Rosen two-step methodology, is plagued by the well-known issue of endogeneity when regressing price on quantity. More recently, Bajari and Benkard introduced their preference inversion approach, which arrives at heterogenous preferences without the second-stage regression and therefore avoids the endogeneity problems. However, it relies on strict functional form assumptions, restricting the MWTP to be a horizontal line for each individual.

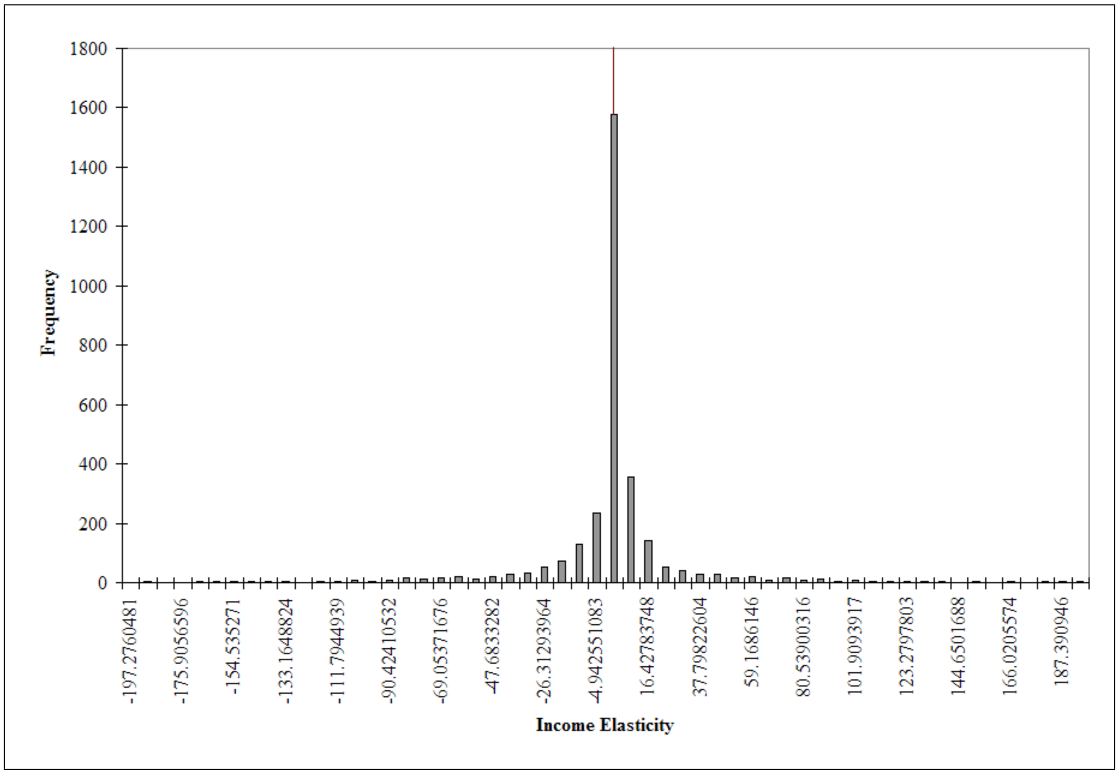

Figure 2.10: Distribution of Income Elasticities using Three-Purchase Panel

By using a preference inversion technique with a panel dataset describing multiple housing purchases per individual, we are able to both avoid the endogeneity problems associated with Rosen and allow MWTP to vary with the level of days exceeding the state 1-hour ozone limit in California. In a first stage set of local-linear regressions, we estimate a separate hedonic gradient for each level of observed pollution in each year of the panel of Bay Area house sales, while controlling for house fixed-effects. In a second stage, we use a panel constructed of individuals who purchase twice over the sample period of 1990 to 2004 to solve for a separate slope and intercept of the MWTP function for each individual. We then use a panel constructed of individuals who purchase three times to additionally solve for a coefficient on non-housing expenditure.

We find that a significant amount of heterogeneity exists, with support for preference-

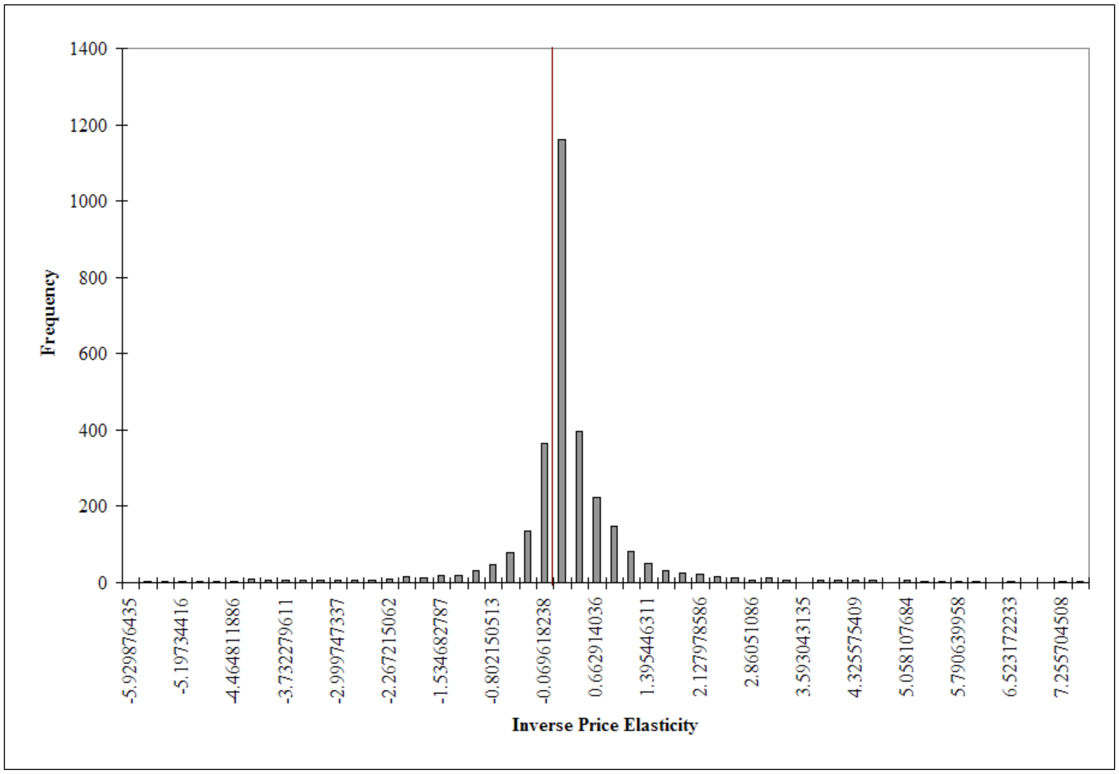

Figure 2.11: Distribution of Inverse Price Elasticities using Three-Purchase Panel

based sorting on ozone. Thus, the use of average MWTP would under-estimate the benefits of reducing ozone pollution in the worst areas.

However, our methodology is not without cost. We start by making very strict assumptions on the stability of preferences over time in our two-purchase model described in Section (2.2.3), *e.g.*, preferences cannot vary with time-varying attributes such as income or number of children. We then show how using data from additional purchases can be used to incorporate these time-varying variables. Thus, to include both income and number of children, the econometrician would need to observe a panel of buyers on at least four purchase occasions. Unfortunately, panel data describing both individual attributes and housing purchases are extremely rare and, if they exist, are not publicly available and often expensive. This critique serves as a motivation for the following chapter, in which we propose a new approach that combines elements of the Bajari and Benkard inversion with those of the traditional methodology to arrive at a new estimator that (i) recovers flexible MWTP functions, (ii) allows for observable and unobservable heterogeneity in preferences, and (iii) does not fall victim to the endogeneity problems that typically arise in recovering MWTP.

Chapter 3

Simple, Consistent Estimation of the Marginal Willingness to Pay Function

3.1 Introduction

Over the last thirty years, property value hedonics [Rosen (1974)] has become commonplace in the non-market valuation of environmental amenities, despite a number of well-known methodological problems. While the estimation of Rosen's first-stage hedonic price function may suffer from omitted variables bias, previous work has shown these problems can be avoided by using panel data describing repeat sales of houses. More difficult problems typically arise in the second-stage of Rosen's procedure – recovery of the marginal willingness to pay (MWTP) function. This function is required is one is interested in valuing any sort of non-marginal change.[1] First, problems arise in simply identifying the second stage of Rosen's procedure. In particular, Brown and Rosen (1982) and Mendelsohn (1985) show that the MWTP function is not identified without making arbitrary functional form assumptions, unless one has data from multiple markets. More recently, Heckman, Ekeland, and Nesheim (2004) have demonstrated that these functional form restrictions can be relaxed if the hedonic price function is estimated non-parametrically, even without multi-market data. Second, recovery of the MWTP function suffers from important endogeneity biases that are difficult to correct with instrumental variables procedures [Epple (1987)]. These difficulties have led most researchers to forgo estimating the MWTP

[1]However, the hedonic valuation of a non-marginal change using such a MWTP function typically ignores the potential for equilibrium re-sorting following the change. New research in estimable equilibrium sorting models [see for example Sieg, Smith, Banzhaf, and Walsh (2004), Bayer, McMillan, and Reuben (2004), Timmins (2007), and Kuminoff (2007)] accounts for this equilibrium response, but does so by making additional structural and distributional assumptions not needed by the hedonic approach.

function, severely restricting the sorts of policies they can consider.

In a 2005 paper, Bajari and Benkard propose a "preference inversion" procedure for recovering heterogeneous measures of MWTP that avoids these endogeneity problems. However, their approach imposes unrealistic constraints on the elasticity of MWTP. In particular, it restricts the demand curve for the amenity to either have zero elasticity (*i.e.*, a flat MWTP function) or to have an elasticity of -1.[2] In the previous chapter, we show how data describing repeat purchase decisions by individual home buyers can be used to relax these constraints. However, extending the inversion procedure proposed by Bajari and Benkard to purchaser panel data requires strong restrictions on the stability of individual preferences over time. This presents an important practical problem - when we see an individual purchase a new home, it is only reasonable to assume that something about her circumstances (*e.g.*, number of children, presence/absence of a spouse, income, *etc*) may have changed since the purchase of her previous house. Allowing preferences to vary with these variables requires that we see numerous repeat purchases, making the assumption of preference stability over time even more tenuous. By combining the traditional estimation approach (*i.e.*, using an econometric error to allow for unobservable, time-varying determinants of preferences) with the intuition of the Bajari-Benkard inversion, we can overcome this difficulty while still avoiding the endogeneity problems associated with the traditional Rosen two-step estimator.

In this chapter, we combine many of the insights of the Bajari-Benkard inversion approach with the unobserved heterogeneity allowed by more standard estimation techniques to arrive at a new hedonic methodology that allows for flexible and heterogeneous preferences and avoids the endogeneity problems that plague the traditional Rosen two-stage model. The technique is, moreover, easy to implement and computationally light. Using data on ozone pollution in the Bay Area of California, we implement this estimator and find

[2]This result is shown in the Appendix.

that flexibility in the shape of the MWTP function is important. In particular, the slope of the MWTP function is -145.18, that is, an individual's willingness-to-pay to avoid an additional day of exceedance increases by \$145.18 when confronted with an additional day of pollution. On the other hand, using the traditional Rosen two-step approach, we find MWTP functions with the counter-intuitive slope (*i.e.*, MWTP is decreasing in pollution) indicating endogeneity problems.

3.2 Model

3.2.1 A Non-Parametric Fixed-Effects Approach to Recovering the Hedonic Gradient

The first stage of the model estimates the hedonic price function, which relates the price of house j transacted in period t ($P_{j,t}$) to its attributes, both those that vary over time ($Z_{j,t}$) and those that do not (X_j). The gradient can be estimated in any number of ways, although a simple, linear framework may impose unrealistic restrictions on the equilibrium underlying the hedonic price function. Additionally, concern must be paid to the potential bias caused by omitted variables.[3] Using panel data with repeat-sales and controlling for house fixed-effects avoids the bias caused by time-invariant house characteristics, whether observed or unobserved. Thus, following from the previous chapter, we adopt a non-parametric fixed-effects approach for the estimation of the gradient and begin by writing down a flexible representation of the hedonic price function:

$$P_{j,t} \;=\; f(Z_{j,t}) + \xi_j + \nu_{j,t} \tag{3.1}$$

where $f(\cdot)$ is an unspecified, flexible function of time-varying attributes of house j (or its neighborhood) and ξ_j represents all time invariant attributes (whether they are observed

[3]Bias from omitted variables arises when the regressors are not orthogonal to the regression error. This would be the case if data describing important neighborhood or housing attributes (*e.g.*, distance to city-center or curb-appeal) are not available to the researcher, but those variables are correlated with the attribute of interest (*e.g.*, air quality).

by the researcher or not). ν_j is assumed to be distributed *i.i.d.* with mean zero and constant variance σ_ν^2.

In the previous chapter, we flexibly estimate the gradient $f'(Z_{j,t})$, *i.e.*, $\frac{\partial P_{j,t}}{\partial Z_{j,t}}$, using local linear methods, assuming that a linear relationship holds in the neighborhood around $Z_{j,t}$. Local linear regression relies on simple ordinary least squares estimation techniques. However, rather than estimating a single set of OLS parameters, it recovers a different parameter vector for every point in the space of $Z_{j,t}$, weighting more heavily the information contained in data points that are "close" to $Z_{j,t}$.[4] As such, it has the ability to flexibly approximate almost any continuous function.[5]

However, with few restrictions on the shape of the hedonic gradient, the data often reveal an extremely non-monotonic function over the range of the variable of interest (*e.g.*, air quality). This potentially leads to problems in a second-stage inversion, where the observed hedonic price for an individual is assumed to be associated with a choice that arose out of utility maximization. Practically, this implies that the MWTP should cut the gradient from above at the individual's chosen level of Z. An example of this violation is shown in Figure (3.1) for the case when Z consists of a single variable, *e.g.*, ozone. The depicted MWTP function would be inconsistent with local utility maximization at Z_0.

When the Bajari and Benkard inversion procedure was taken to the dataset in the previous chapter, approximately 40% percent of the observations violated this first order condition. Using the estimation technique in Bajari and Benkard (2005), however, does not lead to a problem of identification – in particular, the MWTP's for the 60% of individuals who satisfy the utility maximization condition are not affected by the 40% who do not.

[4]The definition of "close" is a decision made by the researcher, in her choice of kernel bandwidth. This is discussed in more detail below.

[5]See Fan and Gijbels (1996) for a description of local-linear regression in general and Ullah and Roy (1998) for a description of local-linear with fixed-effects. Bajari and Kahn (2005) estimate a flexible hedonic gradient using local-linear techniques with a cross section of housing transactions.

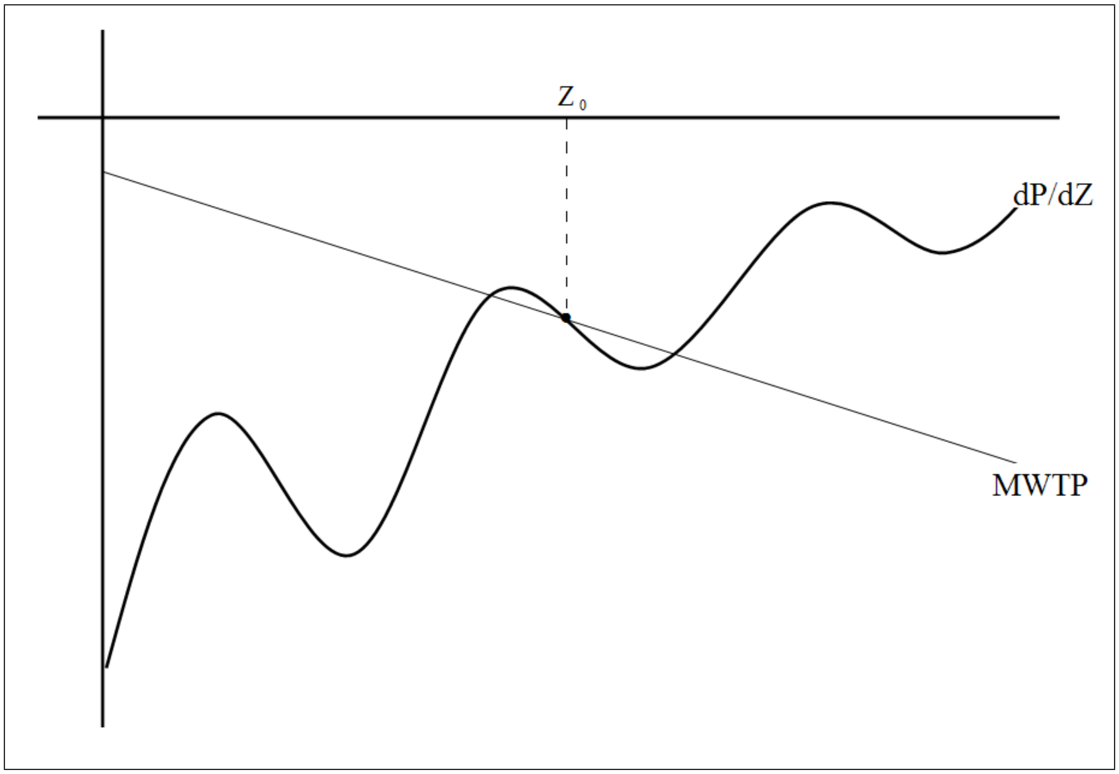

Figure 3.1: Non-Monotonicities Arising from Non-Parametric Gradient Estimation

However, the estimator described below will require that *all* observations are consistent with utility maximizing behavior. To avoid these violations while still allowing for flexibility in the first-stage estimates, we choose to smooth the gradient until these non-monotonicities are no longer a problem in the variable of interest. With a local-linear framework, this would mean moving closer and closer to the globally linear specification for the hedonic price function and a constant hedonic gradient. Thus, in this paper, we flexibly estimate $P_{j,t}(Z_{j,t})$ using local quadratic methods (allowing for the convex shape of the price function and a non-constant gradient, even when over-smoothed), assuming that a quadratic relationship holds in the neighborhood around $Z_{j,t}$. Again this method is straight-forward and relies on simple ordinary least squares estimation techniques.

To set up the estimation, we first take a second-order Taylor series expansion of $f(\cdot)$

around some vector χ (a vector with the same dimension as $Z_{j,t}$):[6]

$$P_{j,t} \;=\; f(\chi) + (Z_{j,t} - \chi)' f'(\chi) + \frac{(Z_{j,t} - \chi)' f''(\chi)(Z_{j,t} - \chi)}{2} + \xi_j + \nu_{j,t} \tag{3.2}$$

Expanding the terms in Equation (3.2) and taking the mean over all sales of house j yields the following:

$$\bar{P}_{j,t} = f(\chi) + (\bar{Z}_j - \chi)' f'(\chi) + \frac{1}{2T_j} \sum_t (Z_{j,t} - \chi)' f''(\chi)(Z_{j,t} - \chi) + \xi_j + \bar{\nu}_{j,t} \tag{3.3}$$

where

$$\bar{P}_j = \frac{1}{T_j} \sum_t P_{j,t} \qquad \bar{Z}_j = \frac{1}{T_j} \sum_t Z_{j,t} \qquad \bar{\xi}_j = \frac{1}{T_j} \sum_t \xi_{j,t} = \xi_j \qquad \bar{\nu}_j = \frac{1}{T_j} \sum_t \nu_{j,t} \tag{3.4}$$

Finally, taking within-house deviations from means allows us to difference away the fixed-effect ξ_j and the function $f(\cdot)$.[7]

$$\begin{aligned} P_{j,t} - \bar{P}_j = {} & (Z_{j,t} - \bar{Z}_j)' f'(\chi) + \\ & \frac{1}{2}(Z_{j,t} - \chi)' f''(\chi)(Z_{j,t} - \chi) - \frac{1}{2T_j} \sum_t (Z_{j,t} - \chi)' f''(\chi)(Z_{j,t} - \chi) + (\nu_{j,t} - \bar{\nu}_j) \end{aligned} \tag{3.5}$$

We denote differences-from-means with "$\sim$" and replace $f'(\cdot)$ with $\beta(\cdot)$.

---

[6]The remainder term associated with the Taylor expansion is ignored.

[7]Losing $f(\cdot)$ from this expression does not pose a problem, as our interest is only in recovering the hedonic gradient, *i.e.*, the slope of the hedonic price function, which is represented non-parametrically by $f'(\cdot)$.

$$\tilde{P}_{j,t} = \tilde{Z}'_{j,t}\beta(\chi) + \frac{1}{2}(Z_{j,t}-\chi)'f''(\chi)(Z_{j,t}-\chi) - \frac{1}{2T_j}\sum_t (Z_{j,t}-\chi)'f''(\chi)(Z_{j,t}-\chi) + \tilde{\nu_{j,t}} \quad (3.6)$$

For the general case where the vector $Z_{j,t}$ is of length k, χ will be of length k, $f'(\chi)$ will be of length k, and $f''(\chi)$ will be a k by k matrix with $\frac{(k^2+k)}{2}$ unique elements. Let $\theta(\cdot)$ be the vector containing the $\frac{(k^2+k)}{2}$ unique elements of $f''(\cdot)$ and let $\tilde{\zeta}$ be a vector containing the unique elements of $\frac{1}{2}(Z_{j,t}-\chi)(Z_{j,t}-\chi)' - \frac{1}{2T_j}\sum_t (Z_{j,t}-\chi)(Z_{j,t}-\chi)'$ (the elements are multiplied by 2 if off the diagonal and 1 if on the diagonal). This allows us to write the problem as the following:

$$\tilde{P}_{j,t} = \tilde{Z}'_{j,t}\beta(\chi) + \tilde{\zeta_{j,t}}(\chi)'\theta(\chi) + \tilde{\nu}_{j,t} \quad (3.7)$$

The parameter vector $[\beta(\chi) \;\; \theta(\chi)]$ can then be recovered with the following minimization procedure:

$$[\beta(\chi) \;\; \theta(\chi)] = \underset{[\beta(\chi) \;\; \theta(\chi)]}{\operatorname{argmin}} \sum_{j=1}^{J}\sum_{t=1}^{T_j}\Big(\tilde{P}_{j,t} - \tilde{Z}'_{j,t}\beta(\chi) - \tilde{\zeta}_{j,t}(\chi)'\theta(\chi)\Big)^2 K_h(Z_{j,t}-\chi) \quad (3.8)$$

where $[\beta(\chi) \;\; \theta(\chi)]$ is evaluated at values of χ equal to all of the observed values of $Z_{j,t}$. We therefore end up with a (potentially) different estimate of the parameter vector for each data point; this is in contrast to a standard OLS estimation of the hedonic gradient, where β is constrained to be the same for every value of $Z_{j,t}$.

Important to the local-quadratic regression procedure is the choice of weights placed on data as one moves further from χ. $K_h(\cdot)$ is the Gaussian kernel:

$$K_h(Z_{j,t}-\chi) = \prod_{k=1}^{\#regressors} \frac{1}{h\hat{\sigma}_{Z_k}}\frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{Z_{k,j,t}-\chi_k}{h\hat{\sigma}_{Z_k}}\right)^2\right\} \quad (3.9)$$

where h represents the kernel bandwidth and $\hat{\sigma}_{Z_k}$ is the standard deviation of the k^{th} element of $Z_{j,t}$.

In our application, $Z_{j,t}$ will be a vector of length 2 and consist of (i) a measure of ozone pollution at house j in year t (*e.g.*, the number of days in the course of each year that the state maximum 1-hour ozone concentration was violated) and (ii) the year of the housing transaction, such that $Z_{j,t} = [ozone_{j,t} \quad year_{j,t}]$.[8] As χ is a (1 by 2) vector in this case, we further specify χ to be comprised of the elements ψ and τ, such that $\chi = [\psi \quad \tau]$. Finally, specifying the following notation:

$$f'(\chi) = [f_\psi(\chi) \quad f_\tau(\chi)] \qquad\qquad f''(\chi) = \begin{bmatrix} f_{\psi\psi}(\chi) & f_{\psi\tau}(\chi) \\ f_{\psi\tau}(\chi) & f_{\tau\tau}(\chi) \end{bmatrix}$$

We estimate the following regression (recovering 5 parameters), assuming that this quadratic relationship holds locally around χ:

$$\begin{aligned}\widetilde{P}_{j,t} \;=\; & \widetilde{ozone}_{j,t} f_\psi(\chi) \;+ \widetilde{year}_{j,t} f_\tau(\chi) \\ + \frac{1}{2}(\widetilde{ozone^2_{j,t}} - 2\psi\widetilde{ozone}_{j,t}) f_{\psi\psi}(\chi) \;& + (\widetilde{ozone_{j,t}year}_{j,t} - \tau\widetilde{year}_{j,t} - \psi\widetilde{ozone}_{j,t}) f_{\psi\tau}(\chi) \\ & + \frac{1}{2}(\widetilde{year^2_{j,t}} - 2\psi\widetilde{year}_{j,t}) f_{\tau\tau}(\chi) \;+ \tilde{\nu}_{j,t} \qquad (3.10)\end{aligned}$$

After visual inspection of the hedonic gradient, we choose our bandwidth for year to be $3 \cdot 1.06 \cdot n^{-\frac{1}{6}}$ where n is the number of observations.[9] For ozone pollution, we choose a bandwidth that sufficiently smooths-out the non-monotonicities that are inconsistent with utility maximization. The choice of bandwidth is therefore sample-dependent.

[8]Including the year that the house transacts controls for the rapid depreciations and appreciations observed over the period of our panel in the Bay Area. Other potential measures of ozone concentration include the 1-hour and 8-hour average maximum concentrations over the course of each year.

[9]Silverman's Rule of Thumb suggests setting the bandwidth value to $1.06 \cdot n^{-\frac{1}{6}}$. We find that, in practice, this tends to under-smooth the estimated hedonic price function. [Silverman (1986)]

Since we allow χ to take on each observed value of the vector Z, the output of this local-quadratic regression procedure provides us with an estimate of $\beta(\chi) = [f_\psi(\chi) \quad f_\tau(\chi)]$ (*i.e.*, the fully flexible slope of the hedonic price function) at each value of Z.[10] We next take this rich information on hedonic prices to the identification of individual preference parameters.

3.2.2 Second Stage Estimation by Hedonic Gradient Inversion

The second-stage of the model begins by writing down the indirect utility function for individual i. Suppose individual i in period t chooses a house to maximize indirect utility over a time-varying amenity such as air quality ($z_{j,t}$) and other (fixed) house attributes (X_j):

$$\begin{aligned} V_{i,j,t} \;=\; & \alpha_{0,i} + \alpha_{1,i,t}X_j + \alpha_{2,i}X_j^2 + \alpha_{3,i,t}z_{j,t} + \alpha_{4,i}z_{j,t}^2 \\ & + \alpha_{5,i}z_{j,t}(I_{i,t} - R_t(X_j, z_{j,t})) + (I_{i,t} - R_t(X_j, z_{j,t})) \quad (3.11) \end{aligned}$$

where $R_t(\cdot)$ is the imputed annual rent or housing expenditure. In practice, we calculate this figure as 7.5% of the observed transaction price.[11] As $R_t(\cdot)$ is a constant fraction of $P_t(\cdot)$, the following relationship holds: $\frac{\partial R_t(\cdot)}{\partial z_{j,t}} = 0.075 \cdot \frac{\partial P_t(X_j, z_{j,t})}{\partial z_{j,t}}$.

As the first-stage of the model provides $\frac{\partial P_t(X_j, z_{j,t})}{\partial z_{j,t}} = f_t'(z_{j,t})$, we can write down the following first order condition for utility maximization:

$$\begin{aligned} & \alpha_{3,i,t} + 2\alpha_{4,i}z_{j,t} + \alpha_{5,i}(I_{i,t} - R_t(X_j, z_{j,t})) \\ & - \alpha_{5,i}z_{j,t} \cdot 0.075 \cdot f_t'(z_{j,t}) - 0.075 \cdot f_t'(z_{j,t}) \;=\; 0 \quad (3.12) \end{aligned}$$

[10]It also provides us with an estimate of $\theta(\chi) = [f_{\psi\psi}(\chi) \quad f_{\psi\tau}(\chi) \quad f_{\tau\tau}(\chi)]$ (*i.e.*, the fully flexible Hessian of the hedonic price function) at each value of Z

[11]This is the commonly used discount in the literature and is estimated in Peiser and Smith (1985).

which can be rewritten as:

$$\alpha_{3,i,t} \; + \; 2\alpha_{4,i} z_{j,t} \; + \; \alpha_{5,i}(I_{i,t} - R_t(X_j, z_{j,t})) \; - \; (\alpha_{5,i} z_{j,t} + 1) \cdot 0.075 \cdot f_t'(z_{j,t}) \;\; = \;\; 0 \quad (3.13)$$

Given data on observable individual attributes (A_i), we specify preference heterogeneity as follows:

$$\begin{aligned} \alpha_{3,i,t} &= \alpha_{3,0} + \alpha_{3,1} A_i + \epsilon_{i,t} \quad (3.14) \\ \alpha_{4,i} &= \alpha_{4,0} + \alpha_{4,1} A_i \\ \alpha_{5,i} &= \alpha_{5,0} + \alpha_{5,1} A_i \end{aligned}$$

where $\epsilon_{i,t}$ is an unobservable determinant of preferences that shifts the intercept of the indirect utility function. This yields the following first-order condition for utility maximization:

$$\alpha_{3,0} + \alpha_{3,1} A_i + \epsilon_{i,t} + 2(\alpha_{4,0} + \alpha_{4,1} A_i) z_{j,t} + (\alpha_{5,0} + \alpha_{5,1} A_i)(I_{i,t} - R_t(X_j, z_{j,t})) - \\ ((\alpha_{5,0} + \alpha_{5,1} A_i) z_{j,t} + 1) \cdot 0.075 \cdot f_t'(z_{j,t}) = 0 \quad (3.15)$$

Until this point, our set-up bears a strong resemblance to the Bajari-Benkard approach. In particular, assuming one has access to a sufficiently long panel of data on home-buyers and is willing to assume that preferences are stable over time, one would implement their approach by evaluating $z_{j,t}$ at individual i's chosen value in period t (i.e., $\rho_{j^*(i),t} = 0.075 \cdot f_t'(z_{j^*(i),t})$) and inverting to recover preferences. This is not possible here because of the unobserved preference heterogeneity term. In contrast, the traditional Rosen (1974) estimation approach uses the unobserved preference heterogeneity term as an econometric error, re-writing equation (3.15) as:

$$\rho_{j^*(i),t} = \frac{\alpha_{3,0} + \alpha_{3,1}A_i + 2(\alpha_{4,0} + \alpha_{4,1}A_i)z_{j,t} + (\alpha_{5,0} + \alpha_{5,1}A_i)(I_{i,t} - R_t(X_j, z_{j,t})) + \epsilon_{i,t}}{(\alpha_{5,0} + \alpha_{5,1}A_i)z_{j,t} + 1} \tag{3.16}$$

In practice, the Rosen (1974) approach requires choosing a functional form for the indirect utility that will yield a linear MWTP curve or just directly imposing a linear functional form for MWTP such as:

$$\rho_{j^*(i),t} = \gamma_{0,0} + \gamma_{0,1}A_i + (\gamma_{1,0} + \gamma_{1,1}A_i)z_{j,t} + (\gamma_{2,0} + \gamma_{2,1}A_i)(I_{i,t} - R_t(X_j, z_{j,t})) + \epsilon_{i,t} \tag{3.17}$$

The standard problem noted by Epple (1987) is that $\epsilon_{i,t}$ and $z_{j^*(i),t}$ will be correlated, and there are not good candidates for strong instruments.[12] This has led most researchers to forgo any attempt at recovering the individual's MWTP function.

Our estimation approach avoids this endogeneity problem and provides a practical method for recovering the MWTP function. We find the vector of utility parameters that maximize the likelihood of observing the values of z actually realized in the data.

The intuition for this approach is summarized in Figure (3.2). Consider a candidate vector of parameters, $\{\alpha_{3,0}, \alpha_{3,1}, \alpha_{4,0}, \alpha_{4,1}, \alpha_{5,0}, \alpha_{5,1}, \sigma\}$, where σ is the standard deviation of $\epsilon_{i,t}$, which is assumed to be distributed i.i.d. $N(0, \sigma^2)$. The mean MWTP function (with $\epsilon_{i,t} = 0$) is plotted for a given level of A_i.[13] The vertical intercept is given by $\bar{C}_i$ and the function is given by Equation (3.16) when $\epsilon_{i,t} = 0$.

[12]Typically, IV approaches have relied on some combination of the following assumptions: (i) certain socio-demographic variables enter directly into the MWTP function while others do not and the excluded variables can be used as instruments for endogenous attribute levels; (ii) the MWTP function only contains socio-demographic variables in linear form, while higher-order terms are excluded from the function and can therefore serve as instruments; or (iii) MWTP does not vary across markets, so that market dummies are excluded and can serve as instruments. In each case, these assumptions are not testable and place arbitrary restrictions on the estimated heterogeneity in MWTP.

[13]In this set-up the MWTP curve is non-linear. However, for simplicity of illustration, we draw a linear MWTP curve. Simpler functional forms for the indirect utility function would yield a linear MWTP curve. In particular, the MWTP curve will be linear if $\alpha_{5,i} = 0$ or if preferences vary with only income (as opposed to non-housing expenditure, which depends on the price of the chosen house).

At these parameter values, we would expect the average individual with attributes A_i to choose a house with attribute $\bar{z}_{j^*(i),t}$. It is now a simple matter to calculate the value of $\epsilon_{i,t}$ for an individual with attributes A_i who chooses a house with attribute $z_{j^*(i),t}$. $\epsilon_{i,t}$ is calculated as the value that moves the intercept (and slope) of the mean MWTP curve such that the adjusted curve passes through the (known point) $f_t'(j^*(i),t)$. In particular, $z_{j^*(i),t}$ implies that the individual's MWTP function intersects the vertical axis at $C_{i,t}$. That is, $\epsilon_{i,t}$ shifts the intercept down by an amount $E_{i,t}$.[14]

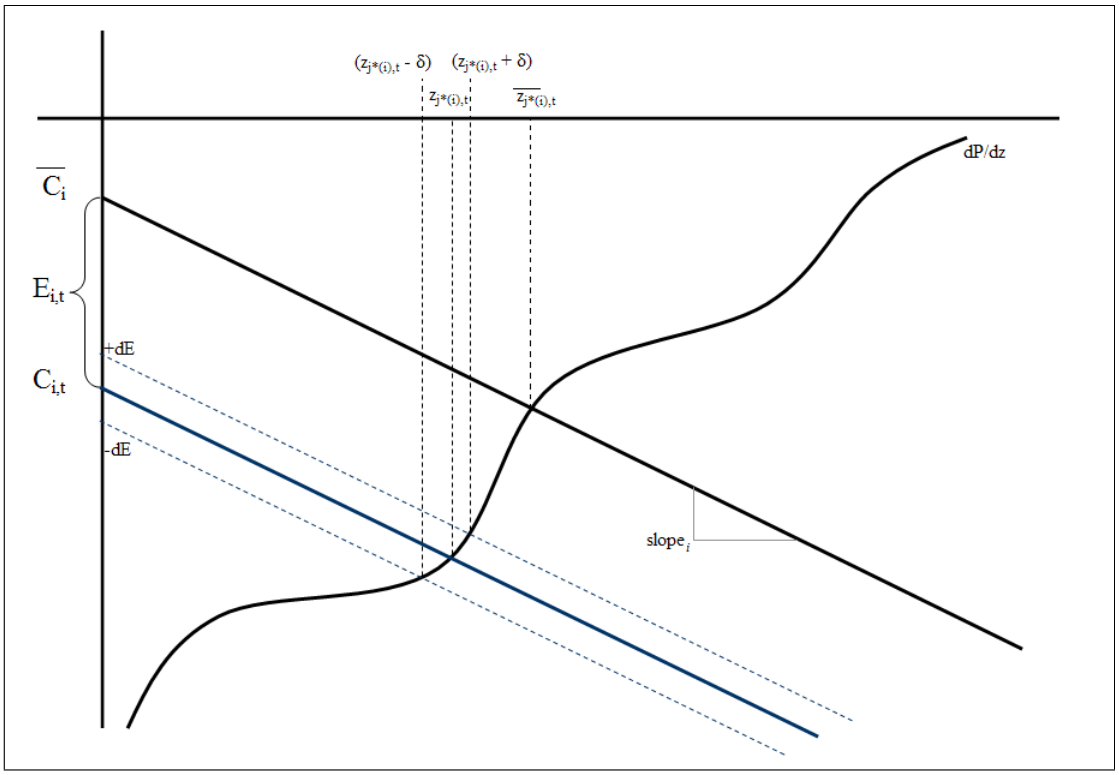

Figure 3.2: Intuition for Hedonic Gradient Inversion

[14]In this set-up $\epsilon_{i,t}$ shifts both slope and intercept. However, in the simple case where the MWTP curve is linear (*i.e.*, $\alpha_{5,i} = 0$), $\epsilon_{i,t}$ shifts only the intercept.

$$\bar{C}_i = \frac{\bar{\alpha}_{3,i}}{1+\bar{\alpha}_{5,i}(I_{i,t}-R_t(X_j,0))} \tag{3.18}$$

$$E_{i,t} = \frac{\epsilon_{i,t}}{1+\bar{\alpha}_{5,i}(I_{i,t}-R_t(X_j,0))}$$

$$C_{i,t} = \frac{\bar{\alpha}_{3,i}}{1+\bar{\alpha}_{5,i}(I_{i,t}-R_t(X_j,0))} + E_{i,t}$$

where

$$\begin{aligned} \bar{\alpha}_{3,i} &= \alpha_{3,0}+\alpha_{3,1}A_i \\ \bar{\alpha}_{4,i} &= \alpha_{4,0}+\alpha_{4,1}A_i \\ \bar{\alpha}_{5,i} &= \alpha_{5,0}+\alpha_{5,1}A_i \end{aligned} \tag{3.19}$$

This implies that individual i's $\epsilon_{i,t}$ is given by:

$$\epsilon_{i,t} = f_t'(z_{j,t})\big((\alpha_{5,0}+\alpha_{5,1}A_i)z_{j,t}+1\big) - \big(\alpha_{3,0}+\alpha_{3,1}A_i+2(\alpha_{4,0}+\alpha_{4,1}A_i)z_{j,t}+(\alpha_{5,0}+\alpha_{5,1}A_i)(I_{i,t}-R_t(X_j,z_{j,t}))\big) \tag{3.20}$$

In order to calculate the likelihood of observing $z_{j^*(i),t}$ given parameters $\{\alpha_{3,0},\alpha_{3,1},\alpha_{4,0},\alpha_{4,1},\alpha_{5,0},\alpha_{5,1},\sigma\}$, we also need to know the absolute value of the derivative associated with the change-of-variables from z to ϵ, $|\frac{\partial\epsilon}{\partial z}|$, at $z_{j^*(i),t}$. This derivative is calculated numerically for $\epsilon^+_{i,t}$ and $\epsilon^-_{i,t}$ at $(z_{j^*(i),t}+\delta)$ and $(z_{j^*(i),t}-\delta)$, respectively. The likelihood of observing $z_{j^*(i),t}$ is then given by:

$$\ell(z_{j^*(i),t}, A_i \; ; \; \alpha_{3,0}, \alpha_{3,1}, \alpha_{4,0}, \alpha_{4,1}, \alpha_{5,0}, \alpha_{5,1}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{1}{2\sigma^2}\epsilon_{i,t}^2\right\}\left|\frac{\partial \epsilon_{i,t}}{\partial z_{j^*(i),t}}\right| \quad (3.21)$$

$$\frac{\partial \epsilon_{i,t}}{\partial z_{j^*(i),t}} = \frac{\epsilon_{i,t}^+ - \epsilon_{i,t}^-}{2\delta}$$

We then find the vector of parameters that maximize the likelihood function:

$$\max_{\alpha_{3,0},\alpha_{3,1},\alpha_{4,0},\alpha_{4,1},\alpha_{5,0},\alpha_{5,1},\sigma} \prod_{t=1}^{T}\prod_{i=1}^{N} \ell(z_{j^*(i),t}, A_i; \alpha_{3,0}, \alpha_{3,1}, \alpha_{4,0}, \alpha_{4,1}, \alpha_{5,0}, \alpha_{5,1}, \sigma) \quad (3.22)$$

As previously discussed, the only complication to this relatively simple procedure is that for a candidate parameter vector to be consistent with utility maximization, it is necessary that the implied MWTP function intersects the hedonic gradient from above at the individual's revealed choice of $z_{j^*(i),t}$. Note that this was also a potential problem in Bajari and Benkard applications and in the panel data application of that procedure described in the previous chapter, and it is a (typically ignored) problem in the canonical Rosen two-step procedure if the hedonic price function is estimated non-parametrically. Were $f_t'(z_{j,t})$ to be estimated with a flexible non-parametric procedure, it would be quite likely that, for some individuals, this utility maximization requirement will not be satisfied. In the application of our estimator, we smooth the data (in our choice of kernel bandwidth) until these non-monotonicites are no longer a problem.

3.3 Data

Implementing our first-stage local quadratic regressions with fixed-effects requires data on multiple sales per house. We use a panel of housing transactions in the Bay Area of California over the period 1990 to 2004. For the second-stage inversion estimator, we require only a cross-section of data that includes both purchase information and individual characteristics. In this case, we assemble a panel dataset by combining information from the real estate transactions dataset and a dataset describing mortgage applicants' demographic characteristics obtained through the Home Mortgage Disclosure Act (HMDA), although repeat observations on purchasers is unnecessary.

3.3.1 Property Transactions Data

The real estate transactions data we employ cover the six core counties of the San Francisco Bay Area (Alameda, Contra Costa, Marin, San Francisco, San Mateo, Santa Clara) over the period 1990 to 2004. The data are purchased from Dataquick and include transaction dates, prices, loan amounts, and buyers', sellers' and lenders' names for all transactions. In addition, the data for the final observed transaction include housing characteristics, such as exact street address, square footage, year built, lot size, number of rooms, number of bathrooms, and number of units in the building.

As our analysis differences out all non-time-varying housing characteristics (and housing attributes are available in the data only for the most recent transaction), it is necessary to ensure we are comparing the same house over time. First, to control for land sales or re-builds, we drop all transactions where "year built" is missing or with a transaction date that is prior to "year built." Second, we want to control for property improvements (*e.g.*, an updated kitchen) or degradations (*e.g.*, water damage) that would not present as a re-build. In a repeat-sales analysis, similar in spirit to Case and Shiller (1989), we regress log prices on a set of house and year dummies (omitting the dummy for the base year, 1990). This gives us a crude measure of yearly appreciation rates in the Bay Area. In years where there is an overall appreciation in housing values (1995-2004), we drop properties that experience a yearly appreciation that is more than five times the average appreciation and properties that experience an average depreciation that is sixty percentage points lower than the average appreciation. In years where there was an overall depreciation in housing values (1990-1994), we drop properties that experience a yearly depreciation that is more than four times higher (in absolute value) than average depreciation and properties that experience an appreciation more than sixty percentage points higher than the average depreciation. In addition, we look at movement within the overall distribution of prices, dropping properties that move more than thirty percentage points in either direction between observed transactions.

Additionally, we drop transactions where the price is missing, negative, or zero. After using the consumer price index to convert all transaction prices into 2000 dollars, we drop one percent of observations from each tail to minimize the effect of outliers. Finally, as we merge in the pollution data using the property's geographic coordinates, we drop properties where latitude and longitude are missing.

Finally, we restrict the sample to properties that sold at least twice. We also drop properties that sell more than once within a given year or more than five times over the fifteen year period. This yields a final sample of 438,492 transactions (*i.e.*, property-year observations), comprised of 191,210 unique properties. Table (3.1) compares the set of repeat sales houses to the full sample of homes in the data. In most dimensions, the samples are very similar.

Table 3.1: Repeat Sale Sample vs. Full Sample

	Full Sample *properties = 594,665*		Repeat-Sale Sample *properties = 191,210*	
variable	mean	median	mean	median
Price (2000 $)	377,684	330,000	377,158	336,673
Sq. Ft. House	1,696.95	1,500.00	1,589.71	1,432.00
Sq. Ft. Lot	9,611.62	5,500.00	8,384.23	5,044.00
Year Built	1966.48	1970	1968.60	1972
Total Rooms	6.38	6	6.21	6
Num. Bedrooms	2.94	3	2.86	3
Num. Bath	2.02	2	1.99	2
Days $\geq$ 0.09 ppm Ozone	2.44	2.09	2.54	2.09

3.3.2 Buyer Characteristics Data

To create a dataset of buyers and their characteristics, we merge the property transactions dataset with the individual attribute data from the HMDA dataset. We use the algorithm found in Bayer, McMillan, Murphy, and Timmins (2008). The match is described in more detail in the previous chapter. This leaves us with a final sample of 52,756 observations comprised of 40,092 individuals.[15]

Table 3.2: Sample of Buyers

n = 40,092		
variable	mean	median
Asian	0.21	0
Black	0.02	0
Hispanic	0.09	0
White	0.59	1
Oth. Race	0.09	0
Income	123,033	98,993
Price	409.943	344.778

3.3.3 Ozone Data

The ozone data we employ are taken from the California Air Resources Board.[16] We use yearly ozone data from all thirty-seven monitors in the nine counties of Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma over the period 1990 to 2004. In particular, we use the monitor data to construct property-specific measures of the number of days exceeding the one-hour state standard (*i.e.*, 0.09 parts per million).

In addition to the ozone readings, the dataset provides information on the "year coverage," or the percent of time (during the relevant high-ozone season) each particular monitor

[15]The algorithm in Bayer *et al.* also matches buyers and sellers through time in the transactions dataset to create a panel dataset that is merged with the HMDA data. For now, we ignore the panel nature of the dataset and treat observations as independent.

[16]Publicly available at www.arb.ca.gov/adam/.

was available and the geographic coordinates of each monitor. Using the coverage variable, we drop monitors with less than 60 percent coverage in a given year (amounting to less than 4 percent of the available monitor-year observations). Using the latitudinal and longitudinal coordinates of the monitors and the properties, we use the "Great Circle" estimator to compute the distance to all monitors from each property. We then create a weighted average for each property of all monitors' readings, using one over distance-squared as the weight. In order to mitigate the boundary effects, we include monitor data from the surrounding counties of Napa, Solano, and Sonoma, in addition to the six counties that appear in our transactions data. Figure (3.3) describes the ozone exceedance data.

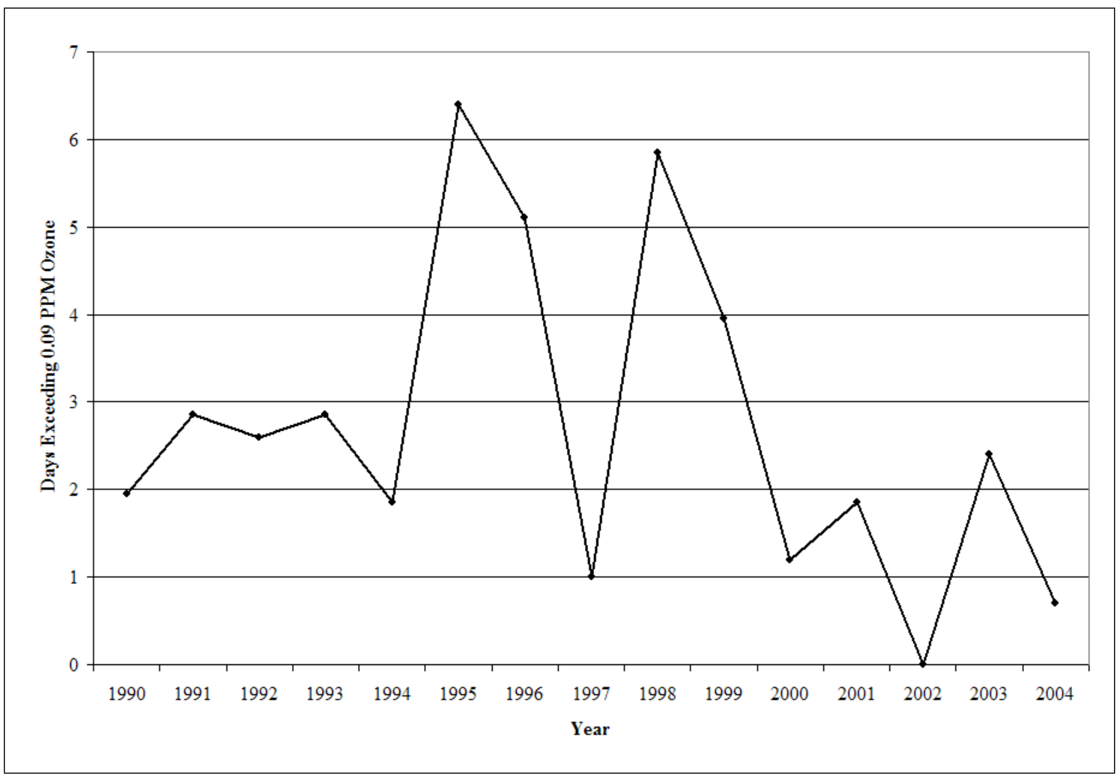

Figure 3.3: Average Number of Days Exceeding CA Ozone Limit

3.4 Results

In our application, we estimate a simplified version of the estimator by restricting the heterogeneity to the following:

$$\begin{aligned} \alpha_{3,i} &= \alpha_{3,0} + \epsilon_{i,t} \\ \alpha_{4,i} &= \alpha_{4,0} \\ \alpha_{5,i} &= \alpha_{5,0} \end{aligned} \tag{3.23}$$

The vector of parameters $\{\alpha_{3,0}, \alpha_{4,0}, \alpha_{5,0}, \sigma\}$ that maximizes the likelihood function is described in Table (3.3).[17]

Table 3.3: Results from the Second-Stage Inversion

parameter		value	standard error	t-statistic
$\alpha_{3,0}$	constant	47.80	0.48	98.81
$\alpha_{4,0}$	ozone exceedance	-72.59	0.30	-241.06
$\alpha_{5,0}$	non-housing expenditure	-1.37E-03	0.12E-04	-111.29
σ_ϵ	std.dev. of the error	455.40	1.78	255.45

Like the Bajari-Benkard inversion method, this estimator generates a distribution of MWTP. Individuals' MWTP are determined by their characteristics (their $z_{j^*(i),t}$ and their $I_{i,t} - R_t(X_{j^*(i)}, z_{j^*(i),t})$) and by their idiosyncratic preference shocks ($\epsilon_{i,t}$). The trimmed distribution of mean MWTP, *i.e.*, MWTP ignoring $\epsilon_{i,t}$, is shown in Figure (3.4).[18] The shape of the distribution is driven by the distribution of individual characteristics.

In addition, like the Rosen second-stage regressions, this estimator allows MWTP to vary with the level of pollution, as well as with non-housing expenditure and time-varying

[17]It is also useful to note that the standard errors presenting in Table (3.3) do not account for the panel nature of the data, *i.e.*, we do not cluster standard errors for the same individual who is observed on multiple purchase occasions.

[18]We trim 1% from each tail to limit the effect of outliers.

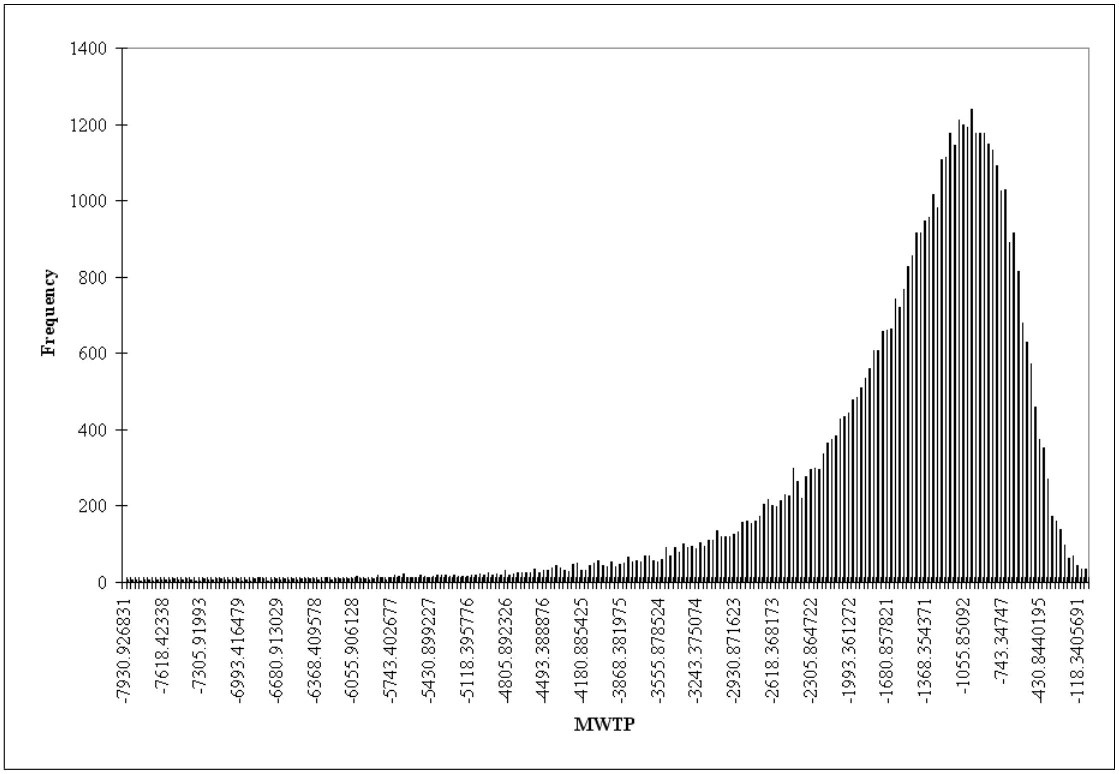

Figure 3.4: Distribution of Mean MWTP

unobservables.

However, unlike the traditional Rosen second-stage, we avoid the endogeneity problems that usually plague these methods. The parameters presented in Table (3.3) are of the expected sign: MWTP is increasing in ozone pollution and non-housing expenditure. The negative coefficient on ozone pollution means that individuals living areas with higher-than-average ozone levels have higher-than-average MWTP to avoid ozone. Therefore, estimating benefits using an average, constant MWTP would underestimate the welfare gains in the dirtiest areas. The negative coefficient on non-housing expenditure means that clean air is a normal good. The slope of the MWTP function is approximately $2 \cdot \alpha_{4,0}$ or -145.18, that is, an individual's MWTP to avoid a day of ozone exceedance increases by \$145.18 with an additional day of exceedance. This relationship holds approximately since $\alpha_{5,0}$ is relatively small.

In Table (3.4) we present the results from the estimation of the traditional, linear Rosen second-stage given by:

$$\rho_{j^*(i),t} = \gamma_0 + \gamma_1 z_{j,t} + \gamma_2 (I_{i,t} - R_t(X_j, z_{j,t})) + \epsilon_{i,t}^{Rosen} \tag{3.24}$$

Table 3.4: Results from the Second-Stage Rosen Estimator

parameter		value	standard error	t-statistic
γ_0	constant	-778.64	3.16	-246.68
γ_1	ozone exceedance	94.24	0.78	121.15
γ_2	non-housing expenditure	-4.13E-05	1.52E-05	-2.72

While the effect of non-housing expenditure on marginal willingness to pay is similar to that found using our estimator, it is clear to see that endogeneity is a problem in the Rosen results when we look at the effect of ozone exceedance. The MWTP function is increasing in ozone pollution, *i.e.*, γ_1 is positive and significant implying that individuals have, on average, a lower willingness to pay in more polluted places. This is equivalent to an upward sloping demand curve for clean air – classic example of the endogeneity bias that arises when $\epsilon_{i,t}^{Rosen}$ and $z_{j,t}$ have a causal relationship. When the hedonic price function is non-linear, buyers simultaneously choose both the hedonic price and the quantity of the environmental amenity that they will consume. This leads to the counter-intuitive sign on ozone pollution.

3.5 Conclusion

In order to measure the welfare consequences of non-marginal changes in amenity values, it becomes necessary to consider individual preference heterogeneity and the possibility of preference-based sorting. However, the standard framework for identifying heterogeneity in preferences, the Rosen two-step methodology, is plagued by the well-known issue of endogeneity when regressing price on quantity. More recently, Bajari and Benkard introduced their preference inversion approach, which arrives at heterogenous preferences without the second-stage regression and therefore avoids the endogeneity problems. However, it relies on strict functional form assumptions, restricting the MWTP to be a horizontal line for each individual. In the previous chapter, we show how panel data can be used to overcome these functional form assumptions. However, this approach requires very strict assumptions on the stability of preferences over time, in addition to a very rich dataset.

In this chapter, we combine many of the insights of the Bajari-Benkard inversion approach with the unobserved heterogeneity allowed by more standard estimation techniques to arrive at a new hedonic methodology that allows for flexible and heterogeneous preferences and avoids the endogeneity problems that plague the traditional Rosen two-stage model. In a first stage of local-quadratic regressions, we estimate a separate hedonic gradient for each level of observed pollution in each year of the panel of Bay Area house sales, while controlling for house fixed-effects. In a second stage, we include a time-varying econometric error and maximize the likelihood of observing each individual's observed choice of ozone pollution.

Using data on ozone pollution (number of days exceeding the California state 1-hour maximum concentration) in the Bay Area, we implement this estimator and find that flexibility in the shape of the MWTP function is important. In particular, the slope of the MWTP function is -145.18, that is, an individual's willingness-to-pay to avoid an additional day of exceedance increases by \$145.18 when confronted with an additional day of pollution. On the other hand, using the traditional Rosen two-step approach, we find MWTP functions with the counter-intuitive slope (*i.e.*, MWTP is decreasing in pollution) indicating endogeneity problems.

Appendix A

Bajari and Benkard Estimator with Cross-sectional Data and Cobb-Douglas Utility

Starting from a Cobb-Douglas specification, utility maximization with respect to non-housing consumption:

$$\max_{(C_i,X_j,z_j)} U_{i,j}(X_j, z_j, C_i) = X_j^{\phi_{X,i}} \; z_j^{\phi_{z,i}} \; C_i \qquad s.t. \qquad C_i + P_j(X_j, z_j) = I_i \tag{A.1}$$

yields the indirect utility function:

$$V_{i,j}(X_j, z_j, I_i) = \phi_{X,i} ln(X_j) \; + \phi_{z,i} ln(z_j) \; + ln(I_i - P_j(X_j, z_j)) \tag{A.2}$$

The first-order condition associated with the optimal housing choice (in particular, the optimal choice of z_j) implies:

$$\frac{\partial V_{i,j}}{\partial z_j} = \frac{\phi_{z,i}}{z_j} - \left(\frac{\partial P_j(X_j, z_j)}{\partial z_j}\right)\left(\frac{1}{I_i - P_j}\right) \; = \; 0 \tag{A.3}$$

$$\phi_{z,i} = z_j \left(\rho_{j^*(i)}\right)\left(\frac{1}{I_i - P_j}\right) \tag{A.4}$$

While this specification does allow MWTP to vary with z_j and income, it does so in a highly restrictive way. In particular, the elasticity of $\rho_{j^*(i)}$ with respect to z_i is restricted to be -1. That such a restriction would be required is not surprising - we are attempting to identify a curve describing an individual's MWTP from only a single data point.

Bibliography

AGUIRREGABIRIA, V., AND P. MIRA (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75(1), 1–54.

ANSELIN, L., AND J. GALLO (2006): "Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects," *Spatial Economic Analysis*, 1(1), 31–52.

ARCIDIACONO, P., AND R. MILLER (2007): "CCP Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity," Mimeo, Duke University.

BAJARI, P., AND C. L. BENKARD (2005): "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 113(6), 1239–1276.

BAJARI, P., C. L. BENKARD, AND J. LEVIN (2007): "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75(5), 1331–1370.

BAJARI, P., AND M. KAHN (2005): "Estimating Housing Demand With An Application to Explaining Racial Segregation in Cities," *Journal of Business and Economic Statistics*, 23(1), 20–33.

BARRO, R., AND X. SALA-I-MARTIN (1992): "Convergence," *The Journal of Political Economy*, 100(2), 223–251.

BARTIK, T. (1987): "The Estimation of Demand Parameters in Hedonic Price Models," *The Journal of Political Economy*, 95(1), 81.

BAYER, P., N. KEOHANE, AND C. TIMMINS (2006): "Migration and Hedonic Valuation: The Case of Air Quality," Mimeo, Duke University.

BAYER, P., S. KHAN, AND C. TIMMINS (2007): "Nonparametric Identification and Estimation in a Generalized Roy Model," Mimeo, Duke University.

BAYER, P., R. MCMILLAN, AND K. RUEBEN (2005): "An Equilibrium Model of Sorting in an Urban Housing Market," Mimeo, Duke University.

BERRY, S. (1994): "Estimating Discrete Choice Models of Product Differation," *Rand Journal of Economics*, 25, 242–262.

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

BLOMQUIST, G., M. BERGER, AND J. HOEHN (1988): "New Estimates of Quality of Life in Urban Areas," *The American Economic Review*, 78(1), 89–107.

Bound, J., D. Jaeger, and R. Baker (1995): "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90(430), 443–450.

Boyle, K., P. Poor, and L. Taylor (1999): "Estimating the Demand for Protecting Freshwater Lakes from Eutrophication," *American Journal of Agricultural Economics*, 81, 1118–1122.

Brown, J., and H. Rosen (1982): "On the Estimation of Structural Hedonic Price Models," *Econometrica*, 50(3), 765–768.

Chattopadhyay, S. (1999): "Estimating the Demand for Air Quality: New Evidence Based on the Chicago Housing Market," *Land Economics*, 75(1), 22–38.

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

Dahl, G. (2002): "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70(6), 2367–2420.

Ekeland, I., J. Heckman, and L. Nesheim (2004): "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 112(S1), 60–109.

Epple, D. (1987): "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," *The Journal of Political Economy*, 95(1), 59.

Epple, D., and H. Sieg (1999): "Estimating Equilibrium Models of Local Jurisdictions," *Journal of Political Economy*, 107, 645–681.

Fan, J., and I. Gijbels (1996): "Local polynomial modelling and its applications, Vol. 66 of Monographs on Statistics and Applied Probability," *London: Chapman Hall,*.

Gemici, A. (2007): "Family Migration and Labor Market Outcomes," Mimeo, University of Pennsylvania.

Gottschalk, P., and R. Moffitt (1994): "The Growth of Earnings Instability in the US Labor Market," *Brookings Papers on Economic Activity*, 1994(2), 217–272.

Hicks, J. (1932): "The Theory of Wages," London: MacMillan Press.

Hoehn, J., M. Berger, and G. Blomquist (1987): "A Hedonic Model of Interregional Wages, Rents, and Amenity Values," *Journal of Regional Science*, 27(4), 605–620.

Hotz, V. J., and R. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–529.

Kennan, J. (2004): "A Note on Approximating Distribution Functions," Mimeo, University of Wisconsin, Madison.

Kennan, J., and J. Walker (2006): "The Effect of Expected Income on Individual Migration Decisions," Mimeo, University of Wisconsin, Madison.

Kuminoff, N. (2007a): "Recovering Preferences from a Dual-Market Locational Equilibrium," Mimeo, Virginia Tech University.

Kuminoff, N. (2007b): "Recovering Preferences from a Dual-Market Locational Equilibrium," Mimeo, Virginia Tech University.

Mendelsohn, R. (1985): "Identifying structural equations with single market data," *Review of Economics and Statistics*, 67(3), 525–29.

Mendelsohn, R., D. Hellerstein, M. Huguenin, R. Unsworth, and R. Brazee (1992): "Measuring Hazardous Waste Damages with Panel Models," *Journal of Environmental Economics and Management*, 22(3), 259–271.

Murray, M. (1975): "The Distribution of Tenant Benefits in Public Housing," *Econometrica*, 43(4), 771–788.

——— (1983): "Mythical Demands and Mythical Supplies for Proper Estimation of Rosens Hedonic Price Model," *Journal of Urban Economics*, 14(3), 326.

Nakamura, A., and M. Nakamura (1998): "Model specification and endogeneity," *Journal of Econometrics*, 83(1-2), 213–237.

Pakes, A., M. Ostrovsky, and S. Berry (2006): "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry/Exit Samples)," NBER Working Paper No. 10506.

Palmquist, R. (1982): "Measuring environmental effects on property values without hedonic regressions," *Journal of Urban Economics*, 11(3), 333–347.

——— (1984): "Estimating the Demand for the Characteristics of Housing," *Review of Economics and Statistics*, 66(3), 394–404.

——— (2003): "Property Value Models," *Handbook of Environmental Economics*, 2.

Parsons, G. (1992): "The effect of coastal land use restrictions on housing prices: A repeat sale analysis.," *Journal of Environmental Economics and Management*, 22(1), 25–37.

Peiser, R., and L. Smith (1985): "Homeownership Returns, Tenure Choice and Inflation," *Real Estate Economics*, 13(4), 343–360.

Pesendorfer, M., and P. Schmidt-Dengler (2003): "Identification and Estimation of Dynamic Games," NBER Working Paper No. 9726.

Roback, J. (1982): "Wages, Rents, and the Quality of Life," *The Journal of Political Economy*, 90(6), 1257–1278.

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

——— (1979): "Wage-Based Indexes of Urban Quality of Life," *Current Issues in Urban Economics, ed. P. Mieszkowski and M. Straszheim*, pp. 74–104, Baltimore, MD: Johns Hopkins University Press.

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3(2), 135–146.

Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55(5), 999–1033.

Savageau, D., and R. D'Agostino (2000): "Places Rated Almanac," CA: IDG Books Worldwide, Inc.

Sieg, H., V. Smith, H. Banzhaf, and R. Walsh (2004): "Estimating the General Equilibrium Benefits of Large Changes in Spatially Delineated Public Goods," *International Economic Review*, 45(4), 1047–1077.

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

Sinnott, R. (1984): "Virtues of the Haversine," *Sky and Telescope*, 68(2), 158.

Sjaastad, L. (1962): "The Costs and Returns of Human Migration," *The Journal of Political Economy*, 70(5), 80–93.

Smith, V. K., and J. Huang (1995): "Can Markets Value Air Quality? A Meta-Analysis of Hedonic Property Value Models," *The Journal of Political Economy*, 103(1), 209–227.

Taylor, L. (2003): "The Hedonic Method," *A Primer on Nonmarket Valuation, ed. Champ, P.A., K.J. Boyle, and T.C. Brown*, pp. 331–393, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Tiebout, C. (1956): "A Pure Theory of Local Government Expenditures," *Journal of Political Economy*, 64(5), 414–424.

Timmins, C. (2007): "If You Can't Take the Heat, Get Out of the Cerrado Recovering the Equilibrium Amenity Cost of Nonmarginal Climate Change in Brazil," *Journal of Regional Science*, 47(1), 1–25.

Timmins, C., and J. Murdock (2007): "A revealed preference approach to the measurement of congestion in travel cost models," *Journal of Environmental Economics and Management*, 53(2), 230–249.

Ullah, A., and N. Roy (1998): "Nonparametric and Semiparametric Econometrics of Panel Data," *Handbook of Applied Economic Statistics, ed. Ullah, A. and D.E.A. Giles*, pp. 579–604, NY, NY: Marcel Dekker.

Biography

Kelly Catherine Bishop was born on July 25, 1978 in Marshfield, Massachusetts. She received her B.A. in Economics (minor in Environmental Science), *cum laude*, from Barnard College, Columbia University in 2000 and her M.A. in Economics, *First Class Honours*, from the National University of Ireland, University College Dublin in 2002. She married Alvin Denis Murphy, a fellow economist, on July 29, 2006. She will graduate from Duke University with a Ph.D. in Economics in 2008 and has accepted an Assistant Professor position at the Olin Business School of Washington University in St. Louis.