

# AN EQUILIBRIUM MODEL OF SORTING IN AN URBAN HOUSING MARKET: A Study of the Causes and Consequences of Residential Segregation

Patrick Bayer  
Department of Economics  
Yale University  
37 Hillhouse Avenue  
New Haven, CT 06511

Robert McMillan  
Department of Economics  
University of Toronto  
150 St. George Street  
Toronto, ON M-5S3G7

Kim Rueben  
Public Policy Institute of California  
500 Washington Street, Suite 800  
San Francisco, CA 94111

October 2002

## Abstract

This paper presents a new equilibrium framework for analyzing economic and policy questions related to the sorting of households within a large metropolitan area. At its heart is a model describing the residential location choices of households that makes explicit the way that individual decisions aggregate to form a housing market equilibrium. The model incorporates choice-specific unobservables, and in the presence of these, a general strategy is provided for identifying household preferences over choice characteristics, including those that depend on household sorting such as neighborhood racial composition. We estimate the model using restricted-access Census data that characterize the precise residential and employment locations of a quarter of a million households in the San Francisco Bay Area, yielding accurate measures of preferences for a wide variety of housing and neighborhood attributes across different types of household. The main economic analysis of the paper uses these estimates in combination with the equilibrium model to explore the causes and consequences of racial segregation in the housing market. Our results indicate that, given the preference structure of households in the Bay Area, the elimination of racial differences in income and wealth would *significantly increase* the residential segregation of each major racial group. Given the relatively small fractions of Asian, Black, and Hispanic households in the Bay Area (each around 10%), the elimination of racial differences in income/wealth (or, education or employment geography) spreads households in these racial groups much more evenly across the income distribution, allowing more racial sorting to occur at all points in the distribution – e.g., leading to the formation of wealthy, segregated Black and Hispanic neighborhoods. The partial equilibrium predictions of the model, which do not account for the fact that neighborhood sociodemographic compositions and prices adjust as part of moving to a new equilibrium, lead to the opposite conclusion, emphasizing the value of the general equilibrium approach developed in the paper. Our analysis also provides evidence that sorting on the basis of race itself (whether driven by preferences directly or discrimination) leads to large reductions in the consumption of public safety and school quality by all Black and Hispanic households, and large reductions in the housing consumption of upper-income Black and Hispanic households.

**Keywords:** Residential Segregation, General Equilibrium, Sorting, Urban Housing Market, Discrete Choice Models, Residential Location Decision, Hedonic Model

# AN EQUILIBRIUM MODEL OF SORTING IN AN URBAN HOUSING MARKET: A Study of the Causes and Consequences of Residential Segregation\*

Patrick Bayer  
Department of Economics  
Yale University

Robert McMillan  
Department of Economics  
University of Toronto

Kim Rueben  
Public Policy Institute of California

October 2002

## Abstract

This paper presents a new equilibrium framework for analyzing economic and policy questions related to the sorting of households within a large metropolitan area. At its heart is a model describing the residential location choices of households that makes explicit the way that individual decisions aggregate to form a housing market equilibrium. The model incorporates choice-specific unobservables, and in the presence of these, a general strategy is provided for identifying household preferences over choice characteristics, including those that depend on household sorting such as neighborhood racial composition. We estimate the model using restricted-access Census data that characterize the precise residential and employment locations of a quarter of a million households in the San Francisco Bay Area, yielding accurate measures of preferences for a wide variety of housing and neighborhood attributes across different types of household. The main economic analysis of the paper uses these estimates in combination with the equilibrium model to explore the causes and consequences of racial segregation in the housing market. Our results indicate that, given the preference structure of households in the Bay Area, the elimination of racial differences in income and wealth would *significantly increase* the residential segregation of each major racial group. Given the relatively small fractions of Asian, Black, and Hispanic households in the Bay Area (each around 10%), the elimination of racial differences in income/wealth (or, education or employment geography) spreads households in these racial groups much more evenly across the income distribution, allowing more racial sorting to occur at all points in the distribution – e.g., leading to the formation of wealthy, segregated Black and Hispanic neighborhoods. The partial equilibrium predictions of the model, which do not account for the fact that neighborhood sociodemographic compositions and prices adjust as part of moving to a new equilibrium, lead to the opposite conclusion, emphasizing the value of the general equilibrium approach developed in the paper. Our analysis also provides evidence that sorting on the basis of race itself (whether driven by preferences directly or discrimination) leads to large reductions in the consumption of public safety and school quality by all Black and Hispanic households, and large reductions in the housing consumption of upper-income Black and Hispanic households.

---

\*We would like to thank Fernando Ferreira (University of California, Berkeley) for outstanding research assistance and Pedro Cerdan and Jackie Chou for help in assembling the data set. We are grateful to Joe Altonji, Pat Bajari, Steve Berry, Gregory Besharov, Greg Crawford, David Cutler, Dennis Epple, James Heckman, Vernon Henderson, Phil Leslie, Enrico Moretti, Robert Moffitt, Tom Nechyba, Steve Ross, Holger Sieg, Kerry Smith, Jon Sonstelie, Chris Taber, Chris Timmins, Chris Udry, Jacob Vigdor, and participants at the meetings of the AEA 2002, Institute for Research on Poverty 2001, and Stanford Institute for Theoretical Economics 2001, and seminars at Chicago, Colorado, Duke, Johns Hopkins, Northwestern, NYU, PPIC, Toronto, UC Berkeley, UC Irvine, UCLA, and Yale for providing many valuable comments and suggestions. This research was conducted at the California Census Research Data Center; our thanks to the CCRDC, and to Ritch Milby in particular. We gratefully acknowledge the financial support for this project provided by the National Science Foundation under grant SES-0137289 and the Public Policy Institute of California.

## 1 INTRODUCTION

A number of important features of the landscape of an urban housing market are determined by the way that households sort among its neighborhoods. This sorting affects residential stratification on the basis of race, income, and other family attributes, the congestion of the transportation network, and the distribution of school quality, crime, property tax bases, and housing prices throughout the urban area. It also has important welfare implications. A full understanding of these implications requires knowledge of the preferences of the heterogeneous households in the metropolitan region and a model that describes how these preferences aggregate to form an equilibrium. The primary goal of this paper is to provide these necessary components.

To that end, we develop an equilibrium model of sorting in an urban housing market and provide a general strategy for identifying preferences in the presence of social interactions in the location decision. Building on McFadden's (1978) discrete choice framework, our model allows households to have preferences for a wide variety of housing and neighborhood attributes, including many that depend explicitly on the way that households sort across neighborhoods in equilibrium (e.g. the quality of local schools, the neighborhood crime rate, and the sociodemographic composition of the neighborhood). Each household's preferences are allowed to vary with its own characteristics, including its wealth, income, education, race, employment location (taken as given), and family composition. The model also provides a well-defined characterization of how these preferences aggregate to determine the equilibrium in an urban housing market; under a set of reasonable assumptions, we demonstrate that a sorting equilibrium always exists in this framework.

The model is estimated using newly available, restricted Census micro-data that provide precise geographic information on the residential locations of a quarter of a million households in the San Francisco Bay Area in 1990. Because the sorting equilibrium is not generically unique, we develop an estimation strategy that permits estimation in the presence of multiple equilibria, exploiting the fact that in any equilibrium, each household chooses its location optimally conditional on the decisions made by the other households in the metropolitan region. This strategy does not require us to compute the equilibrium as part of the estimation procedure, thereby allowing the estimation of hundreds of heterogeneity parameters in a computationally feasible manner.

Following Berry, Levinsohn, and Pakes (1995), we allow explicitly for unobserved differences in the quality of houses and neighborhoods. In so doing, we bring an important endogeneity problem to the forefront of the analysis – namely that the value (or rent) of a house and any other neighborhood attributes determined by the sorting of households are likely to be highly correlated with unobserved house and neighborhood attributes. And we provide a general strategy for identifying the model in the face of this endogeneity problem, developing an appropriate set of instruments for endogenous choice characteristics.

We show that instruments rise naturally out of the logic of the choice model itself. Because each household's location decision is affected by the full set of available alternatives, the housing prices and sociodemographic composition of any particular neighborhood will be partly dependent on the wider availability of choices in the urban housing market. In particular, characteristics of the housing stock and land use in surrounding neighborhoods can be expected to influence, through the market equilibrium, prices and sociodemographics of a given neighborhood. At the same time, as long as the surrounding neighborhoods are sufficiently distant, it is unlikely that their fixed characteristics are correlated with the unobserved features of a given neighborhood that affect household utility, allowing them to serve as valid instruments.

The estimated model yields precise measures of the full set of preference parameters which, along with the characterization of how these preferences aggregate to determine the market equilibrium, can be used to explore a wide variety of economic questions concerning sorting in the urban housing market. The model is particularly useful for carrying out urban policy analysis, providing a way to measure the general equilibrium effects of a policy in terms of its impact on the sociodemographic composition, house values (and rents), school quality, and crime rates of each neighborhood of the metropolitan region, its impact on the intensity of usage of the transportation network, and clear measures of the policy's distributional consequences in terms of income, race, and other household attributes.

#### *Relation to Previous Models of Sorting in an Urban Housing Market*

Our framework draws on two main lines of research in the empirical urban economics literature. Following the seminal work of McFadden (1973, 1978), many researchers have used a discrete choice framework to study residential location decisions, as this framework provides a natural way to estimate heterogeneous preferences for housing and neighborhood attributes.<sup>1</sup> Relative to this literature, the key contribution of our approach is that we explicitly control for the fact that housing prices and neighborhood sociodemographic characteristics are determined as part of the sorting equilibrium, both when estimating the model and conducting counterfactual simulations.<sup>2</sup> In formally characterizing the sorting equilibrium, we build on a vast theoretical literature in urban and public

---

<sup>1</sup> Important applications of this framework can be found in Anas (1982), Anas and Chu (1984), Quigley (1985), and Gabriel and Rosenthal (1989), Nechyba and Strauss (1998), and Duncombe, Robbins, and Wolf (1999).

<sup>2</sup> Developed concurrently with our paper is a closely related study by Bajari and Kahn (2001), which, following Bayer (1999), incorporates error terms that capture the unobserved quality of each location. In their analysis, the authors do not formally model the sorting equilibrium and do not address the correlation between the sociodemographic composition of a community and its unobserved quality, a correlation that is implied by the model.

economics<sup>3</sup> and most directly on the empirical work of Epple and Sieg (1999), which estimates an equilibrium model of community sorting. The key contribution of our framework relative to Epple and Sieg's analysis lies in the flexible form that we adopt for utility, in essence expanding their vertical model of locational differentiation to a more flexible horizontal model of differentiation.<sup>4</sup> By combining what we see as the best features of these two lines of the literature, our goal is to provide a general and flexible framework useful for analyzing a wide range of economic and policy questions in urban economics and local public finance.

### *The Causes and Consequences of Segregation*

The main economic analysis of the paper uses the estimated equilibrium model of sorting to explore the causes and consequences of racial segregation in the housing market. As the seminal work of Thomas Schelling (1969, 1971, 1978) makes clear, a number of distinct microeconomic forces may contribute to an aggregate phenomenon such as segregation. Most obviously, racial segregation could be driven by individual residential choices related to race, either because of direct preferences for the race of one's neighbors or through discrimination in the housing market. The correlation of race with other household characteristics that influence residential sorting - income, wealth, language, immigration experience, and education - could also give rise to a sizeable amount of segregation if these other characteristics are important in shaping residential location decisions. As Schelling noted, "color is correlated with income, and income with residence; so even if residential choices were color-blind and unconstrained by organized discrimination, whites and blacks would not be randomly distributed across residences" (page 144, Schelling (1971)). Other basic mechanisms such as shared social networks or across-race differences in preferences for housing or neighborhood attributes may also contribute to observed segregation patterns. Our equilibrium model of sorting allows us to account for a variety of these potential causes of segregation explicitly.

We begin our analysis of racial segregation by using the equilibrium model to better understand the forces underlying the observed level of racial segregation. In addition to distinguishing the causes of segregation, we also provide evidence on a potentially important consequence of segregation that arises because the single residential location decision simultaneously determines consumption of housing, commuting, and a wide variety of local goods

---

<sup>3</sup> Important contributions to this literature date back to the work of Tiebout (1956) and include the work of Epple, Filimon, and Romer (1984, 1993), Benabou (1993, 1996), Fernandez and Rogerson (1996), and Nechyba (1997, 1999) among others.

<sup>4</sup> In practice, the vertical model constrains households with different characteristics and income to make the same trade-offs between community characteristics so that workers employed in the suburbs, for example, are restricted to have the same preferences for central residential locations relative to other community characteristics as workers employed in the central city. The problem of considering preferences for neighborhood sociodemographics is complicated within the Epple and Sieg framework by the fact that preferences for these characteristics may differ quite non-monotonically across households of different races and ethnicities. Simply including them as part of a public goods index would place undue constraints on these preferences. Epple and Sieg (1999) do not include such

(including neighborhood racial composition). In the presence of this bundled consumption decision, strong preferences in any dimension (e.g., for neighborhood racial composition) distort consumption in other dimensions, especially when the available set of housing options is limited in some important way. In the presence of segregating preferences, it may be difficult for a household to simultaneously satisfy its preferences for neighborhood racial composition and other local goods when the number of households of the same race is relatively small and particularly when the household has significantly different preferences from the majority of households of the same race. The second goal of our analysis of racial segregation is to shed light on this issue, examining the extent to which racial interactions in the location decision accentuate differences in the consumption of housing, school quality, and public safety between white households and those of other races.

#### *Relation to Previous Segregation Literature*

Our analysis departs from most of the prior segregation literature in both its focus and methodology. Much of the prior literature has been concerned with documenting segregation patterns, particularly between black and white households, and how these have changed over time.<sup>5</sup> Recent studies that explore the extent to which segregation might be driven by the correlation of race with other household characteristics include Borjas (1998) and Bayer, McMillan, and Rueben (2002). Both papers examine how the propensity of households to live in segregated neighborhoods varies with other household attributes, including income, education, language, and immigration experience, providing an indication of the extent to which these other household characteristics affect segregation. In forming exact predictions as to how the observed segregation patterns would change if the correlation of race and other household attributes were altered, these studies necessarily condition on features of the urban housing market that are not likely to be primitives. The predictions of the equilibrium model, in contrast, are built on more reasonable primitives of the urban housing market – the underlying distribution of choices in the urban area and preferences across different types of household.

While we do not attempt to make such a distinction in this paper, a number of studies have focused on distinguishing whether segregation arises because of centralized discriminatory practices or the decentralized residential location decisions made by the households of a metropolitan area, each with preferences defined over the race of their neighbors. These studies have typically used data characterizing differences in the prices paid for comparable houses by households of different races to distinguish whether segregation is decentralized sorting on the basis of preferences or discrimination. These studies have focused exclusively on race-based explanations for

---

measures in their analysis, which may seriously bias estimates of preferences for local public goods, as these sociodemographic characteristics are likely to be highly correlated with the observed local public goods.

<sup>5</sup> See Massey and Denton (1987, 1989, 1993), Miller and Quigley (1990), and Harsman and Quigley (1995), for instance.

segregation.<sup>6</sup> The consequences of segregation have also been explored in another body of research that assesses, for example, how across-MSA differences in the degree of segregation affect important outcomes such as educational attainment and wages.<sup>7</sup> None of these papers, however, examines the effect of segregation on racial gaps in the consumption of housing and local public goods.

#### *Data and a Preview of Results*

Our analysis is facilitated by access to newly available restricted-access Census data, as mentioned above. Unlike publicly available Census data, which match each household with a PUMA (a Census area of at least 100,000 residents), these provide a household's residential and employment locations at the level of a Census block (a Census area with approximately 100 residents), allowing us to characterize each household's actual neighborhood much more accurately than has been possible in past studies. The Census data also provide us with detailed information on the households in the sample, including each household member's race, education, income, age, immigration status, employment status, and job location. Using these new Census data as a centerpiece, we have assembled an extensive data set characterizing the housing market in the San Francisco Bay Area. This combines housing and neighborhood sociodemographic data drawn from the Census with neighborhood-level data on schools, air quality, climate, crime, topography, geology, land use, and urban density.

The estimated model provides the most complete picture of the preferences of the households in a major metropolitan region to appear in the literature to date. We obtain precise estimates of the mean valuations across all households of a variety of house and neighborhood attributes, including attributes determined by the way households sort across neighborhoods. The latter include the racial composition of neighborhoods by households of different levels of wealth and education. We also obtain a series of estimates showing how preferences across these choice characteristics vary with household characteristics. In particular, our estimates of racial interactions indicate that there is a strong tendency of households of a given race to be willing to pay much more to live in neighborhood with households of the same race.<sup>8</sup>

The main economic analysis of the paper uses this estimated preference structure along with the equilibrium model to calculate the new sorting equilibrium that arises as the result of a change in the model's

---

<sup>6</sup> Notable papers in this line of research include King and Mieszkowski (1973), Schnare (1976), Yinger (1978), Schafer (1979), Follain and Malpezzi (1981), Chambers (1992), Kiel and Zabel (1996), and Cutler, Glaeser, and Vigdor (1999). Perhaps the most definitive study is by Cutler, Glaeser, and Vigdor (1999), which examines segregation patterns over the full course of the 20<sup>th</sup> century, concluding that centralized racism was much more important in driving segregation in the earlier part of the century.

<sup>7</sup> See Borjas (1995) and Cutler and Glaeser (1997) for important contributions.

<sup>8</sup> It is important to stress that we cannot distinguish whether the estimated racial interactions in the residential location decision are due to the preferences of each race for living with neighbors of the same race or to discrimination in the housing market. We discuss this issue at greater length in Section 5.1 below.

primitives. To explore the causes and consequences of racial segregation, we conduct counterfactuals that eliminate racial differences in income, education, and employment locations as well as experiments that eliminate the preferences that give rise to social interactions in the residential location decision – for instance, preferences for living with households of the same and other races. Our results indicate that the elimination of racial differences in income and wealth (or education) would lead to a *significant increase* in the segregation of each major racial group in the Bay Area given the preferences of the current residents. This result and others associated with eliminating racial difference in education and the geographic distribution of employments leads to one of the fundamental conclusions of our analysis: given the relatively small fractions of Asian, Black, and Hispanic households in the Bay Area (each around 10%), the elimination of racial differences in income/wealth (or, education or employment geography) spreads households in these racial groups much more evenly across the income distribution, allowing more racial sorting to occur at all points in the distribution – e.g., leading to the formation of wealthy, segregated Black and Hispanic neighborhoods. The partial equilibrium predictions of the model, which do not account for the fact that neighborhood sociodemographic compositions and prices adjust as part of moving to a new equilibrium, lead to the opposite conclusion, emphasizing the value of the general equilibrium approach developed in the paper.

Our analysis also provides evidence that sorting on the basis of race itself (whether driven by preferences directly or discrimination) leads to large reductions in the consumption of public safety and school quality by all Black and Hispanic households, and large reductions in the housing consumption of upper-income Black and Hispanic households.<sup>9</sup> When the portion of the preference structure that generates racial interactions in the location decision is eliminated, upper-income Black and Hispanic households in particular are much more likely to choose owner-occupied housing, larger houses, and neighborhoods with much higher levels of school quality and public safety, neighborhoods that also have a much higher fraction of other high-income and white neighbors. These results therefore point to a fundamental consequence of racial sorting in the housing market – namely, a distortion in the consumption of housing and local public goods by members (especially wealthy members) of racial groups with a small numbers of individuals in parts of the income distribution.

The remainder of this paper is organized as follows: In Section 2, we set out the modeling framework and describe the equilibrium properties of the model. The extensive new data set that we have assembled for the analysis is described in Section 3, and estimation of the model is discussed in Section 4. Here, we also relate our model to other methods of estimating willingness-to-pay measures for house and neighborhood attributes. Section 5 discusses issues of identification and interpretation that arise in our sorting model. The next two sections of the

---

<sup>9</sup> As noted previously, we remain agnostic throughout this paper as to whether these interactions arise as the result of the preferences of each race for living with neighbors of the same race or discrimination in the housing market. While this distinction has important welfare implications, the point made here concerning the impact of racial



paper present our empirical analysis: the parameter estimates of the model are given in Section 6, and Section 7 characterizes the pattern of racial segregation in the Bay Area, before setting out results from our general equilibrium simulations. Section 8 concludes.

## 2 AN EQUILIBRIUM MODEL OF SORTING IN THE URBAN HOUSING MARKET

We begin our analysis by setting out an equilibrium model of the housing market, first describing the central component of this model - a discrete choice framework that governs each household's residential location decision - before developing the equilibrium properties of the model.

### 2.1 The Residential Location Decision

The residential location decisions of all households in the San Francisco Bay Area are modeled as a discrete choice of a single residence. The utility function specification is based on the random utility model developed in McFadden (1978) and the specification of Berry, Levinsohn, and Pakes (1995), which includes choice-specific unobservable characteristics.

In the model, each household chooses its residence  $h$  to maximize its utility, which depends on the observable and unobservable characteristics of its choice. Let  $X_h$  represent the observable characteristics of house  $h$  other than price that vary with the household's housing choice and let  $p_h$  denote its price. The observable characteristics of a housing choice include characteristics of the house itself (e.g., size, age, and type), its tenure status (rented vs. owned), and the characteristics of its neighborhood (e.g., sociodemographic composition, school, crime, topography, and air quality). Household  $i$ 's optimization problem is given by:

$$(2.1) \quad \underset{(h)}{\text{Max}} \quad V_h^i = \mathbf{a}_X^i X_h - \mathbf{a}_D^i D_h^i - \mathbf{a}_p^i p_h + \mathbf{x}_h + \mathbf{e}_h^i$$

where  $\mathbf{x}_h$  is the unobserved quality of each housing unit, including the unobserved quality of the corresponding neighborhood. The  $\mathbf{a}_D^i D_h^i$  term in the utility function captures the disutility of commuting – the negative impact of the distance between household  $i$ 's workplace and house  $h$ . The final term of the utility function,  $\mathbf{e}_h^i$ , is an idiosyncratic error term that captures unobserved variation in household  $i$ 's preference for a particular housing choice.

Each household's valuation of choice characteristics is allowed to vary with its own characteristics,  $Z_i$ , including education, income, race, employment status, and household composition. We also assume that each working household is initially endowed with a primary employment location,  $l_i$ . We treat employment status and

---

interactions on the consumption of local public goods by a population with relatively small numbers remains

employment location as exogenous variables throughout this paper.<sup>10</sup> Each parameter associated with housing characteristics, distance to work, and price,  $\mathbf{a}_j^i$ , for  $j \in \{X, D, p\}$ , is allowed to vary with a household's own characteristics,

$$(2.2) \quad \mathbf{a}_j^i = \mathbf{a}_{0j} + \sum_{r=1}^R \mathbf{a}_{rj} Z_r^i,$$

so equation (2.2) describes household  $i$ 's preference for choice characteristic  $j$ . The first term captures the taste for the choice characteristic that is common to all households and the other terms capture observable variation in the valuation of these choice characteristics across households with different socioeconomic characteristics. This heterogeneous coefficients specification allows for great variation in preferences across different types of household.<sup>11</sup>

The specification of utility given in equations (2.1)-(2.2) contains two stochastic components that allow the model flexibility in explaining the observed data. The first component is the house-specific unobservable,  $\mathbf{x}_i$ . This term captures the common value of unobserved (to the econometrician) aspects of a particular house and its neighborhood, that is, value shared by all households. Because many housing and neighborhood attributes are likely to be unobserved in any data set, specifications of the utility function that do not include such unobserved characteristics are likely to lead to biased parameter estimates. The houses in neighborhoods with high levels of unobserved quality, for example, will generally command higher prices and attract higher income households, *ceteris paribus*. Thus analyses that do not account for unobserved characteristics will tend to attribute their impact on utility to observed characteristics with which they are correlated.

The second stochastic component of the utility function is the idiosyncratic term  $\epsilon_h^i$ , which is assumed to be additively separable from the rest of the utility function. We assume that it is distributed according to the Weibull distribution, giving rise to the multinomial logit model. With this assumption, the probability that household  $i$  selects house  $h$ ,  $P_h^i$ , is given by the expression:

$$(2.3) \quad P_h^i = \frac{\exp(\mathbf{a}_X^i X_h - \mathbf{a}_D^i D_h^i - \mathbf{a}_p^i p_h + \mathbf{x}_h)}{\sum_k \exp(\mathbf{a}_X^i X_k - \mathbf{a}_D^i D_k^i - \mathbf{a}_p^i p_k + \mathbf{x}_k)}$$

where  $k$  indexes all possible house choices.

---

regardless of which explanation prevails.

<sup>10</sup> We discuss the impact of these assumptions on the parameter estimates in Section 5 below.

The multinomial logit assumption implies that the ratio of the probabilities between any two choices is independent of the characteristics of the remaining set of alternatives – the IIA property. This property is usually thought to be undesirable, as conveyed by the well-known ‘red bus-blue bus’ example.<sup>12</sup> In housing markets, however, the IIA property helps capture a key feature that is difficult to model directly: the fact that the houses on the market at any time may be thin relative to the full housing stock. Given that a household is limited to purchasing houses that are on the market at the time of search, an increase in the stock of a certain type of housing may significantly increase a household’s probability of choosing that type of house, and perhaps even in a way that resembles the substantial increase generally implied under the multinomial logit assumption.

Two additional elements of the specification given in equations (2.1)-(2.2) limit the impact of the IIA property on the substitution patterns implied by this model. First, the inclusion of the commuting distance term in the utility function ensures that a household is more likely to substitute among choices located near its place of work, giving rise to reasonable substitution patterns in geographic space. Second, the heterogeneous coefficients specification shown in equation (2.2) ensures that while the IIA property holds at the individual level, it does not hold in the aggregate, allowing the model specified in equations (2.1)-(2.2) to give rise to more plausible aggregate substitution patterns. If highly educated households, for example, have a particularly strong taste for school quality, the introduction of a new house in a high quality school district will tend to attract highly educated households, thereby drawing demand away from other houses in high quality school districts. Similarly, houses that are located near each other in geographic space will also tend to be relatively close substitutes in the aggregate, so that the introduction of a new neighboring house will tend to be attractive to those working nearby - the same set of households who presumably found the initial houses attractive in the first place.

## 2.2 Equilibrium: Definition and Properties

While the random utility specification developed above is flexible from an empirical point of view, it also has a convenient theoretical interpretation. Without the idiosyncratic error component,  $\epsilon_h^i$ , this specification would suggest that two households with identical characteristics and employment locations would make identical location decisions. Since this is unlikely to be true in the data, a useful interpretation of  $\epsilon_h^i$  is that it captures unobserved heterogeneity in preferences across otherwise identical households. Thus for a set of households with a given set of observed characteristics, the model predicts not a single choice but a probability distribution over the set of housing

---

<sup>11</sup> While it would also be possible to include random coefficients, i.e., a stochastic term in the preference specification of equation (2.2), which would allowed for unobserved heterogeneity in tastes for each house and neighborhood characteristics, we do not include stochastic terms in the analysis presented in this paper.

<sup>12</sup> In this example, the introduction of an additional though redundant choice takes probabilities away evenly from existing choices, leaving the ratio of probabilities among existing choices unchanged, even though one such choice may be a far closer substitute for the ‘new’ choice than others.

choices. By working with these choice probabilities rather than the discrete decision observed for each household in the sample, it is straightforward to define and explore the properties of a sorting equilibrium for the class of models depicted in equations (2.1)-(2.2). Throughout our analysis, we assume that each household's vector of idiosyncratic preferences  $\mathbf{e}^i$  is observable to all of the other households in the model and we use a Nash equilibrium concept.<sup>13</sup>

Given the household's problem described in equations (2.1)-(2.2), household  $i$  chooses house  $h$  if the utility that it gets from this choice exceeds the utility that it gets from all other possible house choices - that is, when:

$$(2.4) \quad V_h^i > V_k^i \quad \Rightarrow \quad W_h^i + \mathbf{e}_h^i > W_k^i + \mathbf{e}_k^i \quad \Rightarrow \quad \mathbf{e}_h^i - \mathbf{e}_k^i > W_k^i - W_h^i \quad \forall \quad k \neq h$$

where  $W_h^i$  includes all of the non-idiosyncratic components of the utility function  $V_h^i$ . As the inequalities depicted in (2.4) imply, the probability that a household chooses any particular choice depends in general on the characteristics of the full set of possible house choices. In this way, the probability  $P_h^i$  that household  $i$  chooses house  $h$  can be written as a function of the full vectors of house characteristics (both observed and unobserved) and prices  $\{\mathbf{X}, \mathbf{p}, \mathbf{x}\}$ :

$$(2.5) \quad P_h^i = f_h(Z^i, \mathbf{X}, \mathbf{p}, \mathbf{x})$$

as well as the household's own characteristics  $Z^i$ .<sup>14</sup>

When the set of draws  $\{\mathbf{e}_h^i\}$  for each household observed in the data is interpreted as idiosyncratic heterogeneity in preferences for each house, working with choice probabilities is equivalent to assuming that each household that we observe in our sample represents a continuum of households with the same observable characteristics. The choice probabilities depict the distribution of location decisions that would result for a continuum of households with a given set of observed characteristics as each household responds to its particular idiosyncratic preferences. Let the measure of the continuum of households be  $\mathbf{m}$ . This assumption concerning the distribution of households requires a similar assumption about the set of housing choices observed in the sample. In order to make the model coherent, therefore, we also assume that each house observed in the sample represents a continuum of identical houses, and that this continuum also has measure  $\mathbf{m}$ .

---

<sup>13</sup> It is important to point out that other interpretations concerning the exact nature of the idiosyncratic preferences are possible within this framework. We could, for example, treat each household's idiosyncratic preferences as private information and drop the assumption that each household observed in the data stands in for a continuum of other households. In developing the theoretical properties of our model and the estimator, however, we work with the single, consistent interpretation of  $\mathbf{e}$  specified here, attempting to point out in footnotes when other assumptions would be equally valid.

<sup>14</sup> For simplicity of exposition, we have included the household's employment location in  $Z^i$  and the location of the house in  $X_h$ . Note also that the  $h$  subscript on the function  $f$  simply indicates that we are solving for the probability that household  $i$  chooses house  $h$  not that the form of the function itself varies with  $h$ .

### Market Clearing Conditions

Aggregating the probabilities in equation (2.5) over all households yields the predicted number of households that choose each house  $h$ ,  $\hat{N}_h$ :

$$(2.6) \quad \hat{N}_h = \mathbf{m} \bullet \sum_i P_h^i$$

where again  $\mathbf{m}$  represents the measure of the continuum of households with the same observable characteristics as household  $i$ . In order for the housing market to clear, the number of households choosing each house  $h$  must equal the measure of the continuum of houses that each observed house represents:<sup>15</sup>

$$(2.7) \quad \hat{N}_h = \mathbf{m}, \quad \forall h \quad \Rightarrow \quad \sum_i P_h^i = 1, \quad \forall h$$

It is a straightforward extension of the central proof in Berry (1994) to show that under a simple set of assumptions, a unique vector of housing prices clears the market. In particular, we can state the following proposition:

**Proposition 2.1:** If  $U_h^i$  is a decreasing, linear function of  $p_h$  for all households and  $\mathbf{e}$  is drawn from a continuous distribution, a unique vector of housing prices (up to a scaleable constant) solves the system of equations depicted in (2.7), conditional on a set of households  $\mathbf{Z}$  and houses  $\mathbf{X}, \mathbf{x}$ . *Proof:* See Technical Appendix.

Building on Proposition 2.1, the following lemma is also useful for characterizing the properties of a sorting equilibrium in the housing market:

**Lemma 2.1:** If in addition to the assumptions specified in Proposition 2.1,  $U_h^i$  is continuous in a house characteristic  $x_h$  for each household  $i$ , the unique vector of housing prices that clears the market is continuous in  $\mathbf{x}$ . *Proof:* See Technical Appendix.

In proving Proposition 2.1, we show that it is possible to write the solution to (2.7) as a contraction mapping in  $\mathbf{p}$ .<sup>16</sup>

Thus, starting from any vector  $\mathbf{p}$ , an iterative process that increases the prices of houses with excess demand and decreases the prices of houses with excess supply at each iteration leads ultimately to an even spread of households

---

<sup>15</sup> Note that the measure  $\mathbf{m}$  drops out of the market-clearing condition depicted in equation (2.7) and, consequently, simply serves as a rhetorical device for understanding the use of the continuous choice probabilities shown in equation (2.5) in defining equilibrium rather than the actual discrete choices of the individuals observed in the data.

<sup>16</sup> The conditions stated in Proposition 2.1 provide sufficient but not necessary conditions for the existence of a unique vector of market clearing prices. For example, while reasonable, the condition that  $p_h$  enters  $U_h^i$  in a negative manner for *every* household is much more stringent than is actually necessary to ensure the uniqueness result. Ensuring that it is possible to write the solution to the system of equations depicted in (2.8) as a contraction in  $\mathbf{p}$  is as important in practice as proving this system of equations has a unique solution. It is this feature that makes it possible to solve for the unique vector of prices conditional on a set of house and household characteristics in a computationally feasible way.

across houses. Writing this market-clearing vector of prices as  $\mathbf{p}^*(\mathbf{Z}, \mathbf{X}, \mathbf{x})$ , the probability that household  $i$  chooses house  $h$  can be written:

$$(2.8) \quad P_h^i = f_h(Z^i, \mathbf{X}, \mathbf{p}^*(\mathbf{Z}, \mathbf{X}, \hat{\mathbf{1}}), \hat{\mathbf{1}})$$

where the notation  $\mathbf{p}^*(\mathbf{Z}, \mathbf{X}, \mathbf{x})$  indicates that the set of market-clearing prices is generally a function of the full matrices of the household  $\mathbf{Z}$  and house and neighborhood characteristics  $\{\mathbf{X}, \mathbf{x}\}$  that are treated as the primitives of the sorting model.

If the entire set of house and neighborhood characteristics that households value were not affected by the sorting of households across residences, a sorting equilibrium would simply be defined as the set of choice probabilities in equation (2.8) along with the vector of market clearing prices,  $\mathbf{p}^*$ . In this case, since a unique set of prices clears the housing market, the sorting equilibrium would also be unique.

#### *Defining a Sorting Equilibrium with Social Interactions*

For the analysis undertaken in this paper, however, we allow households to have preferences for the sociodemographic characteristics of their neighbors. Such preferences may arise through multiple channels as households may value the characteristics of their neighbors directly and also value other neighborhood attributes such as public safety and school quality that are influenced by neighborhood sociodemographic characteristics. In general, the sociodemographic composition of neighborhood  $n(h)$  can be written in terms of the probability that each household observed in the data chooses each house in that neighborhood. Thus the contribution to the sociodemographic composition of neighborhood  $n(h)$  made by household  $j$  is given by:

$$(2.9) \quad Z_{n(h)}^j = \sum_{k \in n(h)} Z^k \cdot P_k^j$$

and the sociodemographic composition of neighborhood  $n(h)$  can be characterized by the vector of these individual components:  $\mathbf{Z}_{n(h)}$ .

If household  $i$ 's utility from choosing house  $h$  depends explicitly on a function of the sociodemographic characteristics of the occupants of other houses in the same neighborhood  $n(h)$ ,  $g(\mathbf{Z}_{n(h)})$ ,<sup>17</sup> we can write the choice probability defined in equation (2.8) as an explicit function of this function of neighborhood sociodemographic characteristics:

---

<sup>17</sup> For expositional simplicity, we assume that  $g(\mathbf{Z}_{n(h)})$  captures both the direct and indirect channels through which neighborhood sociodemographic characteristics affect utility just described. Note, however, that this function does

$$(2.10) \quad P_h^i = f_h(\mathbf{g}(\mathbf{Z}_{n(h)}), Z^i, \mathbf{X}, \mathbf{p}^*, \hat{\mathbf{i}})$$

Having made the non-price social interactions explicit in the sorting model, we are in a position to define an equilibrium. In particular, a *sorting equilibrium* is defined as a set of choice probabilities  $\{P_h^{i*}\}$  and a vector of housing prices  $\mathbf{p}^*$  such that the following two conditions hold:

- i. The housing market clears according to equation (2.7).
- ii. The set of choice probabilities  $\{P_h^{i*}\}$  is a fixed point of the mapping defined in (2.10), where  $\mathbf{g}(\mathbf{Z}_{n(h)})$  is formed by explicit aggregation of  $P_k^{j*} \forall (j, k)$  according to equation (2.9).

The second condition in this definition ensures that, in equilibrium, each household makes its optimal location decision given the location decisions of all other households.<sup>18</sup>

### Existence

While the equilibrium is defined in terms of the set of optimal household choices and market clearing conditions, it is easier to prove that an equilibrium exists by transforming the problem into a fixed-point problem in the vector of neighborhood sociodemographic characteristics  $\mathbf{g}(\mathbf{Z}_{n(h)})$ . By rewriting equation (2.9) as:

$$(2.11) \quad Z_{n(h)}^j = \dot{\mathbf{a}}_{k\bar{\mathbf{I}}n(h)}^{Z^j} \cdot P_k^j = \dot{\mathbf{a}}_{k\bar{\mathbf{I}}n(h)}^{Z^j} \cdot f_k(\mathbf{g}(\mathbf{Z}_{n(h)}), Z^i, \mathbf{X}, \mathbf{p}^*(\mathbf{g}, \mathbf{Z}, \mathbf{X}, \hat{\mathbf{i}}), \hat{\mathbf{i}})$$

it is easy to see that since  $\mathbf{g}$  is defined over the vector  $\mathbf{Z}_{n(h)}$ , the elements of which are given in equation (2.11), this mapping along with the definition of the function  $\mathbf{g}$  implicitly defines  $\mathbf{g}(\mathbf{Z}_{n(h)})$ . Any fixed point of this mapping,  $\mathbf{g}^*$ , is associated with a unique vector of market clearing prices  $\mathbf{p}^*$  and a unique set of choice probabilities  $\{P_h^{i*}\}$  that together satisfy the conditions for a sorting equilibrium. In this way, finding a vector of prices  $\mathbf{p}^*$  and choice probabilities  $\{P_h^{i*}\}$  that give rise to a sorting equilibrium can be transformed into a fixed-point problem in  $\mathbf{g}(\mathbf{Z}_{n(h)})$ .

We are now able to state the following proposition concerning the existence of an equilibrium:

---

not capture the impact that neighborhood sociodemographic characteristics have on utility through their effect on house price.

<sup>18</sup> Notice that while each household actually makes a discrete location decision, we define the equilibrium in terms of the vector of choice probabilities  $\{P_h^{i*}\}$ . These choice probabilities represent the distribution of location decisions made in equilibrium by the continuum of households that each household  $i$  represents. Note that the alternative assumption that  $\mathbf{e}$  is observed only privately along with a symmetric Bayesian Nash equilibrium concept would allow us to define the equilibrium in terms of discrete location decisions rather than working with the choice probabilities. Existence would continue to hold under this interpretation concerning  $\mathbf{e}$ .

**Proposition 2.2:** If the assumptions of Proposition 2.1 hold, (i)  $U_h^j$  is continuous in  $g(Z_{n(h)})$ , (ii)  $g$  is a continuous function of  $Z_{n(h)}^j$ , and (iii)  $g$  is bounded both above and below, a sorting equilibrium exists. *Proof:* See Technical Appendix.

In the empirical analysis below, we assume that the utility that a household receives from choosing a house is linear in the average sociodemographic characteristics of its neighbors. This assumption ensures that  $U_h^j$  is continuous in  $g(Z_{n(h)})$ ,  $g(Z_{n(h)})$  is a continuous function of  $Z_{n(h)}^j$ , and  $g(Z_{n(h)})$  is bounded by the maximum and minimum values of each household characteristic observed in the data. Thus, if the assumptions of Proposition 2.1 hold, a sorting equilibrium always exists for this class of models.

### *Uniqueness*

While it is straightforward to establish the existence of an equilibrium for the class of models described above, a unique equilibrium need not arise. Consider an extreme example in which two types of households that have strong preferences for living with neighbors of the same type must choose between two otherwise identical neighborhoods. In this case, it is easy to see that the model has multiple equilibria. In particular, two stable equilibria arise with households sorting across neighborhoods by type. When the neighborhoods are identical except for their sociodemographic composition, the matching of each household type with a particular neighborhood is not uniquely determined in equilibrium. Thus, uniqueness is not a generic property of the class of models developed above.

This extreme example, however, gives an unduly pessimistic impression of the likelihood that multiple equilibria arise in this model. Extending the simple example just described, imagine that households of one type have significantly more income than households of the other type, that the quality of one of the neighborhoods is significantly better than that of the other neighborhood in some fixed way, and that households have preferences for neighborhood quality. In this case, while strong preferences to segregate certainly ensure that households again sort across neighborhoods by type, the matching of household type and neighborhood is made much clearer by the marked differences in income and neighborhood quality. In general, a unique equilibrium will arise when the meaningful variation in the exogenous attributes of households, neighborhoods, and houses  $\{Z^i, X_h, \mathbf{x}_h\}$  is sufficiently rich relative to the role that preferences for neighborhood sociodemographic composition play in the location decision.<sup>19</sup>

---

<sup>19</sup> See Bayer and Timmins (2002) for a formal analysis of the conditions under which unique equilibria arise in these models. The discussion here echoes results found earlier in the network and social effects literatures; Katz and Shapiro (1994), for example, write that “consumer heterogeneity and product differentiation tend to limit tipping and sustain multiple networks. If the rival systems have distinct features sought by certain consumers, two or more systems may be able to survive by catering to consumers who care more about product attributes than network size.”



### *Using Choice Probabilities to Define Equilibrium*

In defining a sorting equilibrium, we work with continuous choice probabilities rather than the discrete decisions made by the households observed in the sample. As mentioned, we assume that each household observed in the data represents a continuum of household with identical observable characteristics but distinct idiosyncratic locational preferences. Under this assumption, the sorting equilibrium that arises is not affected by the particular idiosyncratic preferences  $\{\mathbf{e}_h^i\}$  of any single household. The attractiveness of this assumption is obvious as it is the continuity of the choice probabilities that we exploit in proving that a unique vector of prices clears the market and that a sorting equilibrium always exists. If, on the other hand, we interpreted our sample as the literal extent of the housing market, the set of prices that would clear the market (conditional on any finite set of individuals) would no longer be unique.<sup>20</sup> In essence, if an individual had a particularly high draw of  $\mathbf{e}$  for some house, any price high enough to keep everyone else from preferring this house to others in the market and low enough to keep this house as the optimal choice for this individual could work. Despite this range of prices, the existence of an equilibrium would continue as this framework fits within the class of models analyzed by Nechyba (1997, 1999).<sup>21</sup>

As we discuss in Section 4, the same assumptions that allow us to develop the theoretical properties of the model in terms of choice probabilities also play an important role in our estimation strategy. In particular, because uniqueness is not a generic property of the class of models developed above, it is not possible to estimate the model using Maximum Likelihood. We develop instead a GMM estimation procedure that requires that households do not react to the idiosyncratic preferences of any other households in particular. Thus, by ensuring that households can effectively integrate out over  $\mathbf{e}$ , the assumption that we maintain concerning  $\mathbf{e}$  plays an important role in generating a coherent estimation strategy.

---

Likewise, in a closely related model of neighborhood sorting, Nechyba (1999) points out that when “communities are sufficiently different in their inherent desirability, the partition of households into communities is unique.”

<sup>20</sup> This is true as long as  $\mathbf{e}$  continued to be interpreted as individual heterogeneity and each household’s idiosyncratic preferences were common knowledge. An alternative assumption that would generate similar equilibrium properties for our model would be to assume that each household’s idiosyncratic preferences were not common knowledge. This would again ensure that households could not react to the particular idiosyncratic preferences of other households in the market.

<sup>21</sup> It is important to note, however, that given any data set, a researcher would not be able to back out a unique vector  $\mathbf{e}^i$  for each household  $i$  from an observed set of market clearing prices and location decisions. Each household’s equilibrium location decision only reveals that its idiosyncratic preferences for its chosen house exceeded its idiosyncratic preferences for each other house by a certain threshold value. In this way, the *particular* vector  $\mathbf{e}$  for any finite set of households is unidentified, making counterfactual simulations based on calculations of a new equilibrium under alternative assumptions for a *particular* set of households impossible. Knowing the range in which each household’s vector  $\mathbf{e}^i$  must lie, one could conduct counterfactual simulations by randomly drawing a vector  $\mathbf{e}^i$  for each household. This assumption is exactly equivalent to the assumption that we maintain concerning  $\mathbf{e}$  throughout our analysis.

Finally, the use of choice probabilities does not affect the attractive properties of the underlying discrete choice framework related to self-selection. Consider the set of choice probabilities  $P_h^i$  for a particular household observed in the data, which represent the distribution of the discrete decisions made by the continuum of households that the observed household represents. Among this continuum of households, however, those households that choose each particular house  $h$  will be those that get a relatively high draw of  $\epsilon_h^i$  relative to the other houses in the sample. In this way, the set of households predicted to choose each type of house observed in the data are those that place the highest value on it, as governed by both observable household characteristics and idiosyncratic preferences.

### 2.3 A Restricted Version of the Model – A Standard Hedonic Price Regression

Before turning to issues involved with the identification and estimation of the equilibrium model of sorting, it is helpful to examine a restricted version of the model. In particular, consider a specification of the utility function in which all households share the same value for each house except through the idiosyncratic error term:

$$(2.12) \quad U_h^i = \mathbf{a}_{0X} X_h - \mathbf{a}_{0p} p_h + \mathbf{x}_h + \epsilon_h^i$$

Relative to the broader specification described above, this specification eliminates all non-idiosyncratic heterogeneity in preferences and endowments (e.g., employment locations). In this case, the market clearing condition implies that prices adjust so that the mean utility of each alternative is identical and, consequently:

$$(2.13) \quad \mathbf{a}_{0X} X_h - \mathbf{a}_{0p} p_h + \mathbf{x}_h = K \quad \mathbf{P} \quad p_h = \frac{\mathbf{a}_{0X}}{\mathbf{a}_{0p}} X_h + \frac{1}{\mathbf{a}_{0p}} \mathbf{x}_h$$

Equation (2.13) is a standard hedonic price regression. This equivalence makes clear that a hedonic price regression returns the mean valuation of housing and neighborhood attributes when the underlying assumptions of the sorting model specified above (which include the assumption of a fixed stock of housing) are combined with the additional assumption that households have identical preferences for houses and locations.<sup>22</sup>

In the presence of heterogeneity in household preferences for housing and neighborhood characteristics as well as locations, housing units generally provide unequal levels of mean utility in equilibrium. The equilibrium mean utility that a house returns is governed by the relative scarcity of its attributes as well as its location within the urban housing market. Consider, for example, a house with a spectacular view of the Golden Gate Bridge. Such a view is scarce. In this case, we would expect the equilibrium price to reflect the valuation of the view by a very

---

<sup>22</sup> This condition holds no matter what assumption is made concerning the distribution of the idiosyncratic error term and, in fact, holds in the absence of such idiosyncratic preferences.

wealthy individual rather than the mean individual, thereby implying a relatively low level of mean utility in equilibrium. If such a view were less rare, however, the price for such a house would be lower and the level of mean utility higher in equilibrium. Consequently, in the presence of heterogeneous preferences, an adjustment must be made to the price regression of equation (2.13) in order to return mean preferences. As we show in Section 5 below, such an adjustment arises naturally in the course of estimating the equilibrium model.

### 3 DATA

Our analysis is conducted using an extensive new data set built around restricted Census microdata for 1990. These restricted Census data provide detailed individual, household, and housing variables found in the public-use version of the Census, but unlike the public-use data, also include information on the location of individual residences and workplaces at a very disaggregate level. In particular, while the public-use data specify the PUMA (a Census region with approximately 100,000 individuals) in which a household lives, the restricted data specify the Census block (a Census region with approximately 100 individuals). The restricted Census microdata thus allow us to identify the local neighborhood each individual inhabits and to determine the characteristics of that neighborhood far more accurately than has been previously possible with such a large-scale data set.

Our study area consists of six contiguous counties in the San Francisco Bay Area: Alameda, Contra Costa, Marin, San Mateo, San Francisco, and Santa Clara. We focus on this area for three main reasons. First, it is reasonably self-contained. Examination of Bay Area commuting patterns in 1990 reveals that a very small proportion of commutes originating within these six counties ended up at work locations outside the area; and similarly, a relatively small number of commutes to jobs within the six counties originated outside the area. Second, the area contains a racially diverse population. And third, the area is sizeable along a number of dimensions, including over 1,100 Census tracts, and almost 39,500 Census blocks, the smallest unit of aggregation in our data.<sup>23</sup> Our final sample consists of about 650,000 people in just under 244,000 households.

The Census provides a wealth of data on the individuals in the sample – race, age, educational attainment, income from various sources, household size and structure, occupation, and employment location (also provided at the Census block level). Throughout our analysis, we treat the household as the decision-making agent and characterize each household’s race as the race of the ‘householder’ – typically the household’s primary earner. We assign households to one of four mutually exclusive categories of race/ethnicity: Hispanic, non-Hispanic Asian, non-

---

<sup>23</sup> Our sample consists of all households who filled out the long-form of the Census in 1990, approximately 1-in-7 households. In our sample, Census blocks contain an average of 6 households, while Census block groups – the next level of aggregation up – contain an average of 92 households.

Hispanic Black, and non-Hispanic White.<sup>24</sup> To ensure that our sample is representative of the overall Bay Area population, we employ the individual weights given in the Census. Accordingly, 12.3 percent of households are categorized as Asian, 8.8 percent as Black, 11.2 percent as Hispanic, and 67.7 percent of households as White. The full list of the household characteristics used in the analysis, along with means and standard deviations, is given in the upper portion of the first column of the Appendix Table 1.

### *Characterizing Housing Choices*

Households in the model have preferences defined over housing choices, each of which is described by the location of the housing unit,<sup>25</sup> a vector of house characteristics, and a vector of neighborhood characteristics that includes sociodemographic characteristics as well as other information about the neighborhood. The Census data provide a variety of housing characteristics: whether the unit is owned or rented, the corresponding rent or owner-reported value, property tax payment, number of rooms, number of bedrooms, type of structure, and the age of the building.

In constructing neighborhood characteristics, we calculate measures describing the stock of housing in the neighborhood surrounding each house. We also construct neighborhood racial, education and income distributions based on the households within the same block group, a Census region containing approximately 500 housing units.<sup>26</sup> We merge additional data with each house record related to air quality, climate, crime rates, land use, local schools, topography, and urban density. For each of these measures, a detailed description of the process by which the original data were assigned to each house is provided in a Data Construction Appendix.<sup>27</sup> In generating the climate and air quality data at the Census block level, for example, we make use of locally weighted regression techniques to assign data on climate stations and air quality monitoring stations to a lower level of aggregation (in this case, a Census block), as there are far fewer climate stations than Census blocks. The full list of house and

---

<sup>24</sup> The task of characterizing a household's race/ethnicity gives rise to the issue of what to do with mixed race households. One solution would be to assign a household with, for instance, one white and one Hispanic individual a 0.5 measure for both categories while a second option would be to use the characteristics of the household head to define the race/ethnic makeup of the household. We use this second definition and have also omitted the households that do not fit into one of these four primary racial categories (0.7 percent of all households). The results of our analysis are not sensitive to these decisions. Our final sample consists of the 243,350 households that fit into these four racial categories and live in a Census block group that contains at least one other household in our sample.

<sup>25</sup> The latitude and longitude of each house is known at the level of the block, a Census region that contains approximately 100 housing units.

<sup>26</sup> In principle, as we know the location of each house very precisely, neighborhoods could be defined to include all houses within a given radius of the house. In practice, the use of such measures yielded very similar results to those based on conventional Census boundaries, (e.g., Census blocks, block groups, or tracts), and consequently, we use Census block groups when constructing neighborhood sociodemographic measures to facilitate comparison with past research.

<sup>27</sup> This Appendix is available online at [www.economics.utoronto.ca/memillan/dca.htm](http://www.economics.utoronto.ca/memillan/dca.htm).

neighborhood variables, along with means and standard deviations is given in the lower portion of the first column of Appendix Table 1.

### *Employment Access Measures*

Two variables related to employment access are also constructed. First for employed households, we calculate a measure of the distance from the household's principal workplace (defined as the workplace of the individual with the highest labor earnings in the household) to each house in the sample. Here, the location of a house and a job is given by the centroid of the Census block in which each is found, a household's work location also being given in the restricted-access Census data at the block level. Then for every house in the sample, we construct a series of employment access measures based on the local density of jobs that employ households in each of five education categories (<HS, HS Degree, Some College, BA Degree, Advanced Degree). Specifically, for each of these education categories  $E$ , we calculate the access measure  $A_h^E$ , given by:

$$(3.1) \quad A_h^E = \sum_{j \in E} \frac{1}{(d_{jh} + 1)}$$

where  $d_{jh}$  measures the distance from house  $h$  to each job  $j$  in education category  $E$ . This employment access measure for an education category will be large when a house is surrounded by many jobs employing individuals in that education category. Given the education level of each householder, we use the corresponding employment access measure associated with each house to help characterize the quality of the choice.

### *Refining the House Price Variables Provided in Census*

For a variety of reasons, the house price variables reported in the Census are ill-suited for our analysis. House values are self-reported and top-coded, and rents may reflect substantial tenure discounts. Moreover, because we have implicitly defined the model and developed its equilibrium properties in terms of a single price variable for both owner-occupied and rental properties, we must relate house values to rents in some way.<sup>28</sup> Consequently, we make four adjustments to the housing price variables reported in the Census aiming to get a single measure for each

---

<sup>28</sup> This requirement may seem more restrictive than it actually is. Note that we treat ownership status as a fixed feature of a housing unit in the analysis. Thus, whether a household rents or owns is endogenously determined within the model by its house choice. In the model, we allow households to have heterogeneous preferences for home-ownership (a positive interaction between household wealth and ownership, for example, will imply that wealthier households are more likely to own their housing unit, as we find below) and other house characteristics. Moreover, the model could incorporate heterogeneous elasticities of demand for features of a house or neighborhood depending on whether the unit is owned or rented. The use of a single house price variable does not impose any serious restrictions on the model.

unit that reflects what its monthly rent would be at current market prices. We describe the reasoning behind each adjustment here, leaving a detailed description of the methodology for the Data Construction Appendix.<sup>29</sup>

Because house values are self-reported, it is difficult to ascertain whether these prices represent the current market value of the property, especially if the owner purchased the house many years earlier. Fortunately, the Census also contains other information that helps us to examine this issue and correct house values accordingly. In particular, the Census asks owners to report a continuous measure of their annual property tax payment. The rules associated with Proposition 13 imply that the vast majority of property tax payments in California should represent exactly 1 percent of the transaction price of the house at the time the current owner bought the property or the value of the house in 1978. Thus, by combining information about property tax payments and the year that the owner bought the house (also provided in the Census in relatively small ranges), we are able to construct a measure of the rate of appreciation implied by each household's self-reported house value. We use this information to modify house values for those individuals who report values much closer to the original transaction price rather than current market value.

A second deficiency of the house values reported in the Census is that they are top-coded at \$500,000, a top-code that is often binding in California. Again, because the property tax payment variable is continuous and not top-coded, it provides information useful in distinguishing the values of the upper tail of the value distribution.

The third adjustment that we make concerns rents. While rents are presumably not subject to the same degree of misreporting as house values, it is still the case that renters who have occupied a unit for a long period of time generally receive some form of tenure discount. In some cases, this tenure discount may arise from explicit rent control, but implicit tenure discounts generally occur in rental markets even when the property is not subject to formal rent control. In order to get a more accurate measure of the market rent for each rental unit, we utilize a series of locally based hedonic price regressions in order to estimate the discount associated with different durations of tenure in each of over 40 sub-regions within the Bay Area.

Finally, we construct a single price vector for all houses, whether rented or owned. In order to make owner- and renter-occupied housing prices as comparable as possible, we seek to determine the implied current annual rent for the owner-occupied housing units in our sample. Because the implied relationship between house values and current rents depends on expectations about the growth rate of future rents in the market, we estimate a series of hedonic price regressions for each of over 40 sub-regions of the Bay Area housing market. These regressions return an estimate of the ratio of house values to rents for each of these sub-regions and we use these

---

<sup>29</sup> In Section 5, we discuss the implications of using current market prices in estimating the model along with the related issues of moving costs, rent control, and potential lock-in effects associated with Proposition 13.

ratios to convert house values to a measure of current monthly rent. Again, the procedure is described in detail in the Data Construction Appendix.

#### 4 ESTIMATION

Having specified the theoretical framework and described the data, we now present the procedure that we use to estimate the model. We begin by introducing some notation that simplifies the exposition. The terms of the utility function specified in equations (2.1)-(2.2) can be divided into a *choice-specific constant*,  $\mathbf{d}_h$ , an *interaction component*,  $\mathbf{m}_h$ , which includes any parts of the utility function that interact household and choice characteristics, and the *idiosyncratic error term*,  $\mathbf{e}_h^i$ . Thus the utility function can be rewritten as:

$$(4.1) \quad V_h^i = \mathbf{d}_h + \mathbf{m}_h + \mathbf{e}_h^i.$$

where:

$$(4.2) \quad \mathbf{d}_h = \mathbf{a}_{0X} X_h + \mathbf{a}_{0Z} \bar{Z}_h - \mathbf{a}_{0p} p_h + \mathbf{x}_h$$

$$(4.3) \quad \mathbf{m}_h^i = \left( \sum_{r=1}^R \mathbf{a}_{rX} Z_r^i \right) X_h + \left( \sum_{r=1}^R \mathbf{a}_{rZ} Z_r^i \right) \bar{Z}_h - \left( \mathbf{a}_{0D} + \sum_{r=1}^R \mathbf{a}_{rD} Z_r^i \right) p_h^i - \left( \sum_{r=1}^R \mathbf{a}_{rp} Z_r^i \right) p_h$$

In these equations,  $r$  indexes household characteristics, and we have explicitly separated the vector of average neighborhood sociodemographic characteristics  $\bar{Z}$  from  $\mathbf{X}$ . The choice-specific constant  $\mathbf{d}_h$  captures the portion of the utility provided by house  $h$  that is common to all households. In the same way,  $\mathbf{x}_h$ , the unobservable component of  $\mathbf{d}_h$ , captures the portion of unobserved preferences for house  $h$  that is correlated across households, while  $\mathbf{e}_h^i$  represents unobserved idiosyncratic preferences over and above this shared component.<sup>30</sup> Denoting the full set of parameters  $\mathbf{q}$  we subdivide these into two sets in later discussion: the set of interaction parameters in  $\mathbf{m}_h$ ,  $\mathbf{q}_m$  and the set of parameters in  $\mathbf{d}_h$ ,  $\mathbf{q}_d$ . Here, it is worth recalling the assumption from Section 2 that each household  $i$  observed in the sample represents a continuum of otherwise identical households with different idiosyncratic locational preferences and that each house  $h$  observed in the sample represents a continuum of identical houses of the same measure as this continuum of households.

##### 4.1 Estimation without a Generically Unique Equilibrium

---

<sup>30</sup> Another way to describe  $\mathbf{x}_h$  is that it captures the shared portion of the quality of house  $h$  (including the quality of its neighborhood) that is observed by the households in the data but not the econometrician.

Because uniqueness is not a generic feature of the sorting equilibrium, it is clearly not possible to estimate the parameters of the model using Maximum Likelihood. That is, for a set of exogenously given household characteristics  $Z^i$  and house/neighborhood characteristics  $X_h$ , some regions of parameter space give rise to multiple equilibria and therefore do not map uniquely to the set of endogenous variables, which include the matrix of choice probabilities  $\{P^i_h\}$ , the equilibrium vectors of house prices  $\mathbf{p}$ , and neighborhood sociodemographic characteristics  $\bar{\mathbf{Z}}$ . Consequently, we develop a strategy for estimating the parameters using the Generalized Method of Moments (GMM). In this case, the underlying theoretical sorting model need not have a unique equilibrium. Instead, we base the estimation of most of the model's parameters on the assumption that the observed location decisions are individually optimal, given the collective choices made by other households and the vector of market-clearing prices. Our estimation strategy therefore relies on the assumption that an equilibrium in the sorting model exists and is observed, but not the fact that this equilibrium is unique.<sup>31</sup>

In particular, we form moments based on maximizing the probability that each household chooses its observed location conditional not only on  $\mathbf{X}$  but also on  $\mathbf{p}$  and  $\bar{\mathbf{Z}}$ , ignoring the fact that these latter variables are determined as part of the sorting equilibrium. This procedure mirrors an assumption that is typically made when researchers estimate discrete choice models with micro-data – namely that each household is small relative to the whole population and therefore takes all variables as fixed when making its own location decision, even those variables that are endogenously determined. More formally, the validity of our approach derives from our assumption that each household observed in the data represents a continuum of households with distinct idiosyncratic locational preferences. This assumption ensures that households can each effectively integrate out the idiosyncratic preferences of all others when making their own location decisions and consequently that no household's particular idiosyncratic preferences affect the equilibrium. In this way, the vector of idiosyncratic preferences  $\mathbf{e}$  is uncorrelated with the prices and neighborhood sociodemographic characteristics that arise in any equilibrium.

Fitting the observed individual location decisions permits the estimation of the parameters of the interaction term  $\mathbf{ih}$  and the vector of choice-specific constants,  $\mathbf{d}$ . However, the set of observed residential choices provides no information that distinguishes the elements of the choice-specific constant  $\mathbf{d}$ . Consequently, it is necessary to bring additional econometric information to bear on the problem. Given the estimate of  $\mathbf{d}$  obtained from fitting the observed individual location decisions, equation (4.2) is simply a regression equation. The most obvious approach to identifying the parameters of this equation involves forming moments based on covariance restrictions between the observed choice characteristics and  $\mathbf{x}_i$ . It is immediately obvious, however, that forming covariance restrictions

---

<sup>31</sup> Note that this estimation procedure does not require the explicit calculation of an equilibrium, which has the



between  $\mathbf{x}_i$  and  $p_h$ ,  $\bar{Z}_h$ , or any other choice characteristic that depends on neighborhood sociodemographic composition (such as local school quality) is not consistent with the logic of the choice model, as any increase in the unobserved quality of a house typically raises demand for a house and in turn its equilibrium price. Similarly, an increase in the unobserved quality of a neighborhood will tend to increase the price of houses in that neighborhood and alter the sociodemographic composition of the households living there in equilibrium. In estimating equation (4.2), therefore, it is necessary to find a vector of additional instruments  $\mathbf{W}$  for the housing prices and neighborhood sociodemographic characteristics that are determined endogenously in the sorting model. We discuss the specific instruments used in the analysis in Section 5 below.<sup>32</sup>

#### 4.2 The Estimation Procedure

The estimation procedure just outlined is straightforward to implement. For any combination of interaction parameters and house-specific constants,  $\mathbf{d}_i$ , the model predicts the probability that each household  $i$  chooses house  $h$ :

$$(4.4) \quad P_h^i = \frac{\exp(\mathbf{d}_h + \hat{\mathbf{n}}_h^i)}{\sum_k \exp(\mathbf{d}_k + \hat{\mathbf{n}}_k^i)}$$

Maximizing the probability that each household makes its correct housing choice, conditioning on the full set of observed household characteristics  $Z^i$  and choice characteristics  $\{X_h, p_h, \bar{Z}_h\}$ , gives rise to the following log-likelihood function:

$$(4.5) \quad \ell = \sum_i \sum_h I_h^i \ln(P_h^i)$$

where  $I_h^i$  is an indicator variable that equals 1 if household  $i$  chooses house  $h$  in the data and 0 otherwise. The first step of the estimation procedure consists of searching over the interaction parameters and vector of choice-specific

---

attractive effect of significantly reducing the computational burden involved.

<sup>32</sup> In some empirical settings, researchers may not be interested in distinguishing the components of the vector of choice-specific constants,  $\mathbf{d}$ . It is necessary to do so, however, if one is interested in calculating *any* individual's willingness-to-pay for *any* choice characteristic or if one wants to carry out any counterfactuals – or make any predictions – since any systematic changes in location preferences affect  $\mathbf{p}$  and  $\bar{\mathbf{Z}}$  in equilibrium. Distinguishing the components of  $\mathbf{d}$  immediately forces one to address the correlation of  $\mathbf{x}$  with  $\mathbf{p}$  and  $\bar{\mathbf{Z}}$ , giving rise to the need for instruments.

constants to maximize  $\ell$ ,<sup>33</sup> returning estimates of the interaction parameters  $\hat{\mathbf{q}}_m$  and the vector of choice-specific constants  $\hat{\mathbf{d}}$ . The second step of the estimation procedure uses  $\hat{\mathbf{d}}$  along with a set of appropriate instruments to estimate equation (4.2) via instrumental variables.<sup>34</sup>

#### *A Computational Shortcut – Enforcing the Market Clearing Condition*

When the size of the choice set grows large, it becomes infeasible to search freely over the full set of parameters that need to be estimated in the first step of the estimation procedure  $(\delta, \mathbf{q}_m)$ . Conditional on the data and  $\mathbf{q}_m$  however, it is possible to ‘back-out’ an estimate of  $\delta$  by enforcing the market clearing conditions specified in equation (2.7). As it turns out, this procedure returns the vector  $\mathbf{d}$  that maximizes the likelihood function in equation (4.5), conditional on the set of interaction parameters,  $\mathbf{q}_m$ .<sup>35</sup> Operationally, Berry (1994) demonstrates that for any  $\mathbf{q}_m$  a unique vector of choice-specific constants  $\mathbf{d}$  (up to a scaleable constant) satisfies the market-clearing conditions and Berry, Levinsohn, and Pakes (1995) provide a contraction mapping that solves for  $\mathbf{d}$ . For our application, the contraction mapping is simply:

$$(4.6) \quad \mathbf{d}_h^{t+1} = \mathbf{d}_h^t - \ln(\hat{N}_h^t)$$

where  $t$  indexes the iterations of the contraction mapping and  $\hat{N}_h^t$  is the predicted number of households that choose each house. Using this contraction mapping, it is possible to solve quickly for an estimate of the full vector  $\hat{\mathbf{d}}$  even when it contains a large number of elements, and consequently the likelihood function can be concentrated as:  $\ell(\mathbf{d}, \mathbf{q}_m) = \ell^c(\hat{\mathbf{d}}(\mathbf{q}_m), \mathbf{q}_m)$ . This reduces our free parameter search to  $\mathbf{q}_m$  thereby dramatically reducing the computational burden in the first step of the estimation procedure.

Because the issue can lead to confusion, it is important to point out that the vector of choice-specific constants can be estimated even if the number of housing alternatives in the sample is as large as the number of

---

<sup>33</sup> Recall that the likelihood function defined in equations (4.5) conditions on the full set of choice characteristics, including those that are endogenously determined, so that this procedure is not a Maximum Likelihood procedure, despite appearances.

<sup>34</sup> While estimating the two steps of the estimation procedure jointly would increase the efficiency of the estimator if the instruments used in the second stage of the analysis were valid, consecutive estimation ensures that the estimates obtained in the first stage are consistent regardless of the consistency of the instruments used in the second stage. In all of the analysis presented in this paper, we use the two-step procedure.

<sup>35</sup> The derivative of the likelihood function given in equation (4.5) with respect to  $\mathbf{d}_i$  is:

households. In essence, estimation is possible because an increase in any particular  $\mathbf{d}_h$  increases the probability that each household in the sample chooses house  $h$ . This increases the probability that the model correctly predicts the house choice for the household that actually chooses house  $h$ , but decreases the probability that all of the other households in the sample make the correct choice, as is apparent from (4.4). As the logic of the market clearing condition makes clear, the likelihood function is maximized when the model predicts equal demand for each house observed in the sample. In this way, the estimate of  $\mathbf{d}_h$  for each house  $h$  is governed by the total demand for the house rather than solely on the demand of the individual that purchases the house.<sup>36</sup>

### *Using a Random Sample of Alternatives*

Despite the shortcut for estimating the full vector of choice-specific constants introduced in the previous subsection, calculating choice probabilities for each household-house pair at each iteration of the optimization routine quickly becomes computationally infeasible when the number of housing alternatives grows large. Consequently, in order to estimate the model, we employ a sampling framework specified in McFadden (1978) in which a randomly chosen subset of the full set of alternatives is used in calculating each household's contribution to the likelihood function in equation (4.5). Using this sampling framework, McFadden shows that one can obtain consistent estimates of the model's parameters without calculating the full matrix of choice probabilities.

The particular procedure that we use is as follows. We first draw a large sample of households  $S^I$  and their corresponding houses  $S_H$  at random from the full Census data set.<sup>37</sup> This initial draw ensures that  $S^I$  and  $S_H$  are random samples, each representative of the entire San Francisco Bay Area and that the actual house chosen by each household in  $S^I$  is in the full sample of houses  $S_H$ . For each household  $i$  observed in this sample, we then construct a subset  $S_H^i$  of the full set of houses in the Bay Area that consists of the household's chosen house and a random sample of the remaining alternatives in  $S_H$ . In this way, the choice probabilities for each household that are used in the constructing the likelihood function shown in equation (4.5) are given by:

---


$$\frac{\partial \ell}{\partial \mathbf{d}_h} = \sum_{i=h} \frac{\partial \ln(P_h^i)}{\partial \mathbf{d}_h} + \sum_{i \neq h} \frac{\partial \ln(P_h^i)}{\partial \mathbf{d}_h} = \sum_{i=h} (1 - P_h^i) + \sum_{i \neq h} (-P_h^i) = 1 - \sum_i (P_h^i)$$

and, consequently, setting the vector  $\frac{\partial \ell}{\partial \mathbf{d}}$  equal to zero produces the market clearing conditions specified in equation (2.7).

<sup>36</sup> This discussion also makes clear why  $\mathbf{d}$  is subject to a free normalization, as an increase in each element of  $\mathbf{d}$  has no effect on any household's demand for any house.

<sup>37</sup> Census sample weights are used in this step of the analysis, ensuring that our initial sample is representative of the households and houses with the San Francisco Bay Area.

$$(4.7) \quad P_h^i = \frac{\exp(d_h + \hat{m}_h^i)}{\sum_{k \in S_H^i} \exp(d_k + \hat{m}_k^i)}$$

where the sum in the denominator is now taken over only those alternatives in the subset associated with household  $i$ .

In practice, because we estimate choice-specific constants for each house, the precision of the estimation procedure increases greatly if we ensure that each alternative appears in the choice set of the same number of households. To this end, we employ the following random sampling procedure: Starting with the assignment of each household's chosen house, we assign each household a first additional (not chosen) alternative by randomly re-shuffling the full set of houses across households. We then repeat this random re-shuffling of houses as many times as is necessary to generate the desired size of the sample of additional (not chosen) alternatives. In this way, with an additional random draw for each household, we ensure that each alternative is sampled exactly once.

The sampling framework developed by McFadden (1978) also justifies the initial sampling process that generates the full set of households  $S^I$  and houses  $S_H$  used in the analysis (i.e., the use of less than the full census of houses) as long as the IIA property holds at the individual level. That is, if households actually choose from the full census of houses including, for example, those that are not sampled in the long form of the Census, no researcher will observe each household's full choice set in the data. In this case, the assumption that  $d_h$  is distributed according to the Weibull distribution, which gives rise to the IIA property at the individual level, ensures that the econometrician obtains consistent estimates of the model's parameters despite observing only a random sample of the full set of alternatives that each household faces. While other assumptions could be made to justify the use of a sub-sample of the full census of alternatives (i.e., that the observed sample spans the full choice set), the underlying assumption concerning the distribution of  $\mathbf{e}$  justifies the use of the sample without such assumptions.<sup>38</sup>

The use of a random sample of the full census of alternatives for each household necessitates a slight adjustment to the calculation of the predicted number of households that choose each house that is used in the contraction mapping (4.6). In particular, because the sampling procedure ensures that each household's actual choice is included in the subset of alternatives when calculating the choice probabilities shown in equation (4.7), the predicted number of households that chooses each house used in equation (4.6) must be corrected for this inherent over-sampling. This requires the following straightforward adjustment:

---

<sup>38</sup> It is worth noting that this does not prevent a researcher from including other stochastic elements in the utility function such as random coefficients, which allow for unobserved differences across households in willingness-to-pay for choice characteristics. All that is required to justify the general use of this kind of sampling procedure is that the final idiosyncratic component of location preferences be distributed according to the Weibull distribution.

$$(4.8) \quad \hat{N}_h = \frac{(C+1)}{N} \sum_{i=h} P_h^i + \frac{(C+1)(N-1)}{N C} \sum_{h \in S_H^i, i \neq h} P_h^i$$

where  $N$  is the total number of alternatives in the full census,  $C$  is the number of additional (not chosen) alternatives sampled for each household and, consequently, the number of times house  $h$  appears in other household's choice set, and the notation  $i=h$  refers to the household that actually chooses house  $h$ . In equation (4.8), the first term captures the contribution to  $\hat{N}_h$  made by the household who actually chose house  $h$ , while the second term sums the contributions of the other households in the sample which could have chosen house  $h$  (i.e., the house was in the household's randomly drawn choice set) but did not.

If one takes the full set of houses in the metropolitan area to be the relevant choice set,  $N$  is very large, on the order of 1.5 - 2 million for the San Francisco Bay Area. Even if one counts only the number of houses in the full set  $S_H$  used in the analysis,  $N$  is typically large relative to  $C$ , and consequently, equation (4.8) effectively reduces to:<sup>39</sup>

$$(4.9) \quad \hat{N} = \frac{(C+1)}{C} \sum_{h \in S_H^i, i \neq h} P_h^i$$

In calculating  $\hat{N}_h$  for use in backing out the vector of choice-specific constants that ensures that the housing market clears, we essentially do not count the contribution of the household that actually chooses the house. Dropping the choice probability for the household that actually chooses a house from this calculation is intuitive as the sampling framework described above includes this individual-house pair in the analysis in a non-random way.

### *Summary of the Full Estimation Procedure*

The full estimation procedure that we employ can be summarized as follows:

#### 1. Sample and Choice Set Construction

- i. Draw a large sample of households  $S^I$  and their corresponding houses  $S_H$  at random from the full Census data set. The sample of households is used directly in the analysis, while the sample of houses is used in constructing subsets of alternatives for each household in (1.ii).
- ii. For each household  $i$  observed in this sample, construct a subset  $S_H^i$  consisting of the household's chosen house and a random sample of the remaining alternatives in  $S_H$ . These sets are held fixed throughout the remainder of the estimation.

#### 2. Estimation of Interaction Parameters and Choice-Specific Constants

---

<sup>39</sup> For a large enough sample of houses, one might assume that this sample effectively spans the full census of houses in the metropolitan area. In this way, the full census of choices would be represented in the full choice set  $S_H$  drawn initially from the data.

- i. For a given set of interaction parameters (those in  $\boldsymbol{\eta}_h$ ), solve for the vector of choice-specific constants  $\boldsymbol{d}$  that implies that the housing market clears for each house (i.e., that equation (2.7) holds).
- ii. Using the vector of house-specific constants  $\boldsymbol{d}_i$  and  $\boldsymbol{\eta}_h$ , calculate the log-likelihood function given in equation (4.5).
- iii. Search over the interaction parameters until the objective function calculated in (2.ii) is maximized. The estimated choice-specific constants are those calculated in (2.i) at the final iteration.

### 3. Estimation of Choice-Specific Constant Regression

- i. Using the estimated choice-specific constants from (2), estimate equation (4.2) using instrumental variables.

## 4.3 Asymptotic Properties of the Full Estimator

Because our estimation procedure is somewhat non-standard, we discuss briefly the conditions that ensure the consistency and asymptotic normality of our estimates. We begin by clarifying the maintained assumptions concerning the data generating process. The model is estimated on data drawn from a single, large metropolitan area. The complete metropolitan area housing market consists of a total of  $I$  individuals who must choose from  $H$  distinct types of housing, with  $H$  assumed to be less than  $I$ . Each individual  $i$  is characterized by a set of characteristics  $Z^i$  and a set of idiosyncratic preferences  $\{\epsilon^i_h\}$  defined over the full set of distinct housing alternatives, and identically and independently distributed across choices according to the Weibull distribution. Households are assumed to follow the model's decision rule at the true parameter vector. Each distinct housing type  $h$  is characterized by the set of characteristics  $\{X_h, \boldsymbol{\eta}_h\}$ . The  $\{X_h, \boldsymbol{\eta}_h\}$  vectors are assumed to be exchangeable draws from some larger population of possible house types.

We do not observe the full census of households and houses in the metropolitan area, but instead observe a random sample of households  $S^I$  and their corresponding houses  $S_H$  of size  $I^S$  and  $H^S$  respectively. Moreover, the actual house type chosen by each household in  $S^I$  appears in the sample of houses  $S_H$ . Because the sample of houses is drawn randomly from the full sample of houses, the relative market share of each house observed in  $S_H$  is exactly  $k_h/H^S$ , where  $k_h$  is the number of times that a house of type  $h$  is sampled. Thus, within the observed sample of housing alternatives, the relative market share of each house type is known exactly. Finally, the random sampling technique described above is used to draw a subset of housing alternatives for each household  $i$  in sample  $S^I$  that consists of the household's chosen house and a random sample of the remaining alternatives in  $S_H$  of size  $C$ . With this characterization of the data generating process, our problem fits within a class of models for which the asymptotic distribution theory has been developed. In the remainder of this section, therefore, we summarize the requirements necessary for the consistency and asymptotic normality of our estimates and provide some intuition for these conditions.



















































































































