Race, Environment, and Neighborhood Choice

H. Spencer Banzhaf†

Nicholas E. Flores‡

Joshua Sidon ‡

Randall P. Walsh ‡*

May, 2007

Work in Progress

*†Georgia State University; ‡University of Colorado, Boulder. This material is based upon work supported by the National Science Foundation under Grant No. SES-0321566. Any opinions, findings and conclusions or recomendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

1 Introduction

This paper estimates preferences for environmental quality through residential housing choices within the Los Angeles metropolitan area. The work is part of a larger project that investigates issues of environmental justice, *i.e.* disproportionate exposure of minority groups to neighborhood environmental hazards, in a housing equilibrium framework. In particular, the larger project seeks to measure potential differential effects of environmental policy across income and racial groups.

Even if one rejects the notion that minorities or the poor "choose" to remain in more polluted neighborhoods, it is difficult to deny that higher income, mobile white families may move into a minority or poor neighborhood once environmental quality is improved, particularly if neighborhood housing prices are relatively low. Consequently, a policy to improve environmental quality in the neighborhood may result in an influx of higher income whites that are willing and able to afford the new, improved environmental quality. Cities with increasing road congestion such as Denver, Dallas, and Washington D.C. are already experiencing this phenomenon. In response to increasing congestion, areas close to the city center are being redeveloped and selling at premium prices. These once-affordable neighborhoods are quickly becoming too expensive for lower income families. Improving other dimensions of environmental quality such as the removal or regulation of neighborhood point source pollution may have a similar effect.

Understanding preferences for housing, neighborhood environmental quality, and neighborhood racial composition across income and racial groups is necessary to adequately analyze the equilibrium effects of environmental policy with an eye toward environmental justice considerations. Preference estimates from this paper will ultimately be used to conduct environmental policy simulations. These simulations

will particularly focus on the welfare effects of environmental policy across different income and racial groups.

While the goals of the larger project are very specific, implementing the research requires crossing many methodological bridges. This paper contributes to the environmental economics and urban economics literature in several important ways. First, the paper explicitly models the decision whether to rent or to own in a way that provides estimates of housing capitalization rates by income groups. Understanding preferences for renting versus owning is crucial for understanding the implications of large scale policy changes since over 30% of households in the U.S. do not reside in owner occupied housing. Housing value appreciation or depreciation that results from large scale policy changes, such as environmental quality improvements, will accrue to owners while renters face higher rental rates. Without explicit consideration of potential differential policy effects for renters and owners, polices designed to address issues of environmental justice may inadvertently result in welfare losses for those the policies are intended to help. Moreover modeling the tenure decision as an endogenous choice will allow for the simulation of policies that may complement those designed to enhance environmental equity, such as housing subsidies for the poor.

Second, similar to Bayer, McMillan, and Ruben (2004) and Bajari and Kahn (2005), the paper models preferences for neighborhood racial composition. Jointly modeling preferences for environmental quality and racial preferences will provide important insights into issues of environmental justice. Third, the paper develops a general methodology for estimating heterogenous preferences for environmental quality in a discrete/continuous estimation framework from data on housing sales transactions and public use tabulations from the 2000 U.S. Population Census. This methodology is an important contribution because it avoids the use of restricted access Census micro data that is used in Bayer, McMillan, and Rueben (2004) while

facilitating finer spatial resolution than is provided through the Public Use Microdata Series used by Bayer, Keohane, and Timmins (2005) and Bajari and Kahn (2005). Finer spatial resolution provides the opportunity to analyze highly localized environmental issues. Taken together, these contributions serve to expand the applicability and understanding of discrete choice housing models and their relation to measuring environmental quality.

The remainder of the paper is organized as follows. Section 2 provides a brief overview of the modeling approach and its connection to the existing literature. Section 3 presents the choice model and estimation strategy. Section 4 discusses describes the data used in estimation while Section 5 provides implementation details and results.

2 Connection to Existing Literature

The primary modeling objective of this paper is to estimate preferences for environmental quality while allowing for heterogeneity across groups defined by both income and race. As noted above, the focus on measuring preferences across income/race groups provides the opportunity to evaluate the potential differential effects of environmental policy. Households choose a neighborhood location, whether to rent or own, and then choose the amount of housing in that neighborhood. The observed population proportions of an income/race group across neighborhoods facilitates an estimation strategy similar to the approach outlined in Berry (1994) and Berry, Levinsohn, and Pakes (1995).

Indirect utility for a given neighborhood choice relies on a tractable, but flexible functional form that allows for reasonable substitution patterns between housing and non-housing consumption. In particular, the approach does not restrict housing

price and income elasticities to unity. Using the prescribed indirect utility function, discrete choice over neighborhoods is modeled using a linear neighborhood fixed effect discrete choice model. The neighborhood fixed effect incorporates observable and non-observable neighborhood characteristics. The decision to rent or own is explicitly modeled. Distinct neighborhood price indices are developed for rental units and owner occupied housing values. The model estimation provides what are essentially capitalization rates that are specific to groups. This approach is new to the literature in which capitalization is typically inferred from outside of the model. A second stage of estimation decomposes the fixed effects using neighborhood characteristics including environmental quality, racial composition, and school quality. This second estimation stage allows for the use of standard instrumental variables techniques as suggested by Berry (2004) while allowing for unobservable neighborhood characteristics that can vary by group.

This work follows a long line of housing choice papers. Early work on discrete choice modeling of housing choice dates back to the mid 1970's with an application by Quigley (1976) and subsequent theoretical work by McFadden (1978). This early work utilized individual-level data in a standard discrete choice modeling framework in which relevant housing attributes were assumed to be observable or uncorrelated with the random utility error component. Later work specifically modeled unobservable characteristics with an emphasis on empirically testing predictions from Tiebout (1956) type jurisdictional sorting models. This work includes Epple and Sieg (1999), and Epple, Romer, and Sieg (2001). Using very similar models, Sieg, Smith, Banzhaf, and Walsh (2004) and Walsh (2004) estimate housing and consumption preferences in order to simulate the general equilibrium effects of large scale environmental policy changes. Walsh considers open space policies in Wake County, North Carolina while Sieg, Smith, Banzhaf, and Walsh consider Southern California air quality improve-

ments. The work so far mentioned that uses an unobserved characteristics approach emphasizes equilibrium analysis and for this reason imposes a single crossing property on preferences in order to characterize equilibrium conditions. The resulting models facilitate income heterogeneity and environmental quality preference heterogeneity, but this heterogeneity only applies to a single public good. Though these models have been very influential, they do not readily handle multiple neighborhood-specific characteristics such as racial composition and environmental quality, nor do they facilitate heterogeneous preferences by readily identifiable groups.

The work in this paper emphasizes estimating preferences for identifiable income/racial groups and the estimation approach is very closely related to work found in Bayer, Keohane, and Timmins (2005). Bayer, Keohane, and Timmins model choice of U.S. Census Metropolitan Statistical Area (MSA) from the Public Use Microdata Series. Location choice in their model provides housing services, environmental quality as measured by particulate matter, and wage income. Though single neighborhood price indices are estimated using both rental rate information and self reported housing values, they do not explicitly model the decision to rent or own. There are two groups of people in their model, those with some college training or who graduate from college and those without college training. Bayer, Keohane, and Timmins explicitly model the propensity to move as part of the choice framework. They conduct the analysis to estimate willingness to pay to avoid particulate matter and then contrast this estimate with values obtained from standard wage hedonic estimation.

While the analysis in this paper pursues a similar estimation strategy to Bayer, Keohane, and Timmins, there are several important differences. Locations in this paper are much smaller geographically since aggregate location choice data for types is constructed up from the Census block level as opposed to only having information relevant to the much coarser Public Use Microdata Areas. Types in this paper are

defined across income and race, rather than education, and location choices are made within metropolitan area rather than across metropolitan areas. In contrast to the mobility constraints, captured through inclusion of birth location in the model, that are estimated in Bayer, Keohane, and Timmins, this paper uses job location as a mobility factor within the metropolitan area. Finally, this paper considers preferences for neighborhood racial composition which is likely to be important for considering the effect of policies across income/race groups.

Two influential papers model preferences for racial composition. Bajari and Kahn (2005) model housing choices in Atlanta, Chicago, and Dallas using separate models for each area. To this end, they use Pubic Use Microdata. As with Bayer, Keohane, and Timmins (2005), the data only facilitates spatial resolution at the Public Use Microdata Area. The identification strategy in their model uses a three-step estimation procedure developed in Bajari and Benkard (2005), an estimation approach that is very different from the approach in this paper. Bajari and Kahn first estimate a nonparametric hedonic price function. Then they use this estimated price function to recover household-specific marginal valuations for continuous goods in the indirect utility function based on first order conditions implied by utility maximization. In the third stage they infer individual taste coefficients which they allow to vary according to individual demographic characteristics. Their analysis provides for taste heterogeneity by allowing estimated taste parameters to vary linearly with household demographics. As part of their analysis, Bajari and Kahn adjusts prices for owner occupied housing using a 7.5% capitalization rate. Hedonic estimation includes an owner occupied indicator and the third stage provides an estimate of the taste for ownership. In contrast to Bajari and Kahn, this paper does not assume a capitalization rate and instead utilizes separate rental and price indices. Tastes for renting and owning then translate directly into inferred capitalization rates across income/race

groups.

Bayer, McMillan, and Rueben (2004) model housing choice with an emphasis on measuring racial preferences, allowing for heterogeneity across racial groups by allowing taste parameters to vary linearly with household demographic characteristics (similar to Bajari and Kahn). Their analysis emphasizes racial segregation and simulates counterfactuals to explore the impact of increasing minority incomes on housing market segregation. Though not a specific focus of their analysis, preferences for air quality are estimated through their statistical analysis. Bayer, McMillan, and Rueben estimate their model from restricted access Census micro data for the San Francisco Bay area. The restricted access Census micro data provides very detailed information including household specific demographic information including job location, structural information on the house, self-assessed housing value or rental rate, and geographic location which facilitates attaching very refined neighborhood characteristics. Restricted access data accommodates very fine spatial resolution. A single price for rental and owner occupied housing services is constructed using estimated capitalization rates for 40 sub-regions in the San Francisco Bay area. The approach to modeling renting or owning is very similar to Bajari and Kahn. The richness of the data they use facilitates a fully discrete choice approach with no quantity choice after neighborhood choice. The estimation approach in Bayer, McMillan, and Rueben incorporates unobserved characteristics while utilizing an innovative approach to modeling an equilibrium. Their modeling approach uses what looks like a standard discrete choice model, but then imposes the condition that the probability of a particular house in their sample being chosen must equal one across their sample. The probability condition allows them to identify a unit-specific fixed effect that includes neighborhood characteristics. The unit specific fixed effect is analogous to the identification of neighborhood fixed effects employed in this paper as well as

Bayer, Keohane, and Timmins (2005). In contrast to this paper and Bayer, Keohane, and Timmins, the unit-specific fixed effect of Bayer, McMillan, and Rueben does not vary by type, though interactions between neighborhood characteristics and household demographics serve a similar purpose. While the estimation strategy of Bayer, McMillan, and Rueben and the data they use to estimate their model are considerably different than the approach of this paper, the inferred preferences are quite similar in that both models provide estimates of heterogeneous preferences for neighborhood racial composition and environmental quality.

3 Modeling Neighborhood Choice

An individual household i is distinguished by type k, where type is comprised of income class (I classes) x race (R races). Job location is important from the standpoint of implied commute distance, D. In this model there are L different job locations. For individual i, distance from neighborhood j to their job location is given by $D_{i,j}$.

- i indexes the household.
- j indexes the neighborhood.
- k indexes type; there are RxI types.
- $t_i \in \{r, o\}$ indexes individual i's tenure status.
- A_{jk} indexes the capitalization (or annualization) rate in region j for type k.
- l indexes the job location; there are $L \leq J$ job locations.
- $D_{i,j}$ is the distance from their job location to neighborhood j for individual i.

- X_j is a vector of observable neighborhood characteristics including environmental quality, school quality, crime for neighborhood j. These variables are treated as exogenous in them model.

- Z_j is the endogenous racial composition of neighborhood j

- $\xi_{j,k,t}$ is an unobservable Neighborhood X Tenure attribute.

Utility from choosing neighborhood j and tenure status t for household i of class k is determined by the levels of observable neighborhood attributes, X_j and Z_j, unobservable neighborhood attributes $\xi_{j,k,t}$, the quantity of housing consumed once j is chosen, $H_{i,j,k,t}$, a composite commodity $C_{i,j,k,t}$, and a zero mean, random error component, $\varepsilon_{i,j,k,t}$. This error is assumed to be generated from a type 1 extreme value distribution. Household i makes housing and consumption choices subject to the budget constraint $I_k = P_{t,j}H_{i,j,k,t} + C_{i,j,k,t}$ where $P_{t,j}$ is equal to the rental price index for community j, $\rho_{r,j}$, when renting and in the case of owning is equal to the 'capitalization rate' for i's type in location j times the sales (owner occupied) price index for community j, $A_{jk}\rho_{o,j}$. I_k is income for household i of type k. Finally the distance from i's job location to neighborhood j, $D_{i,j}$, impacts utility since more commuting time is required to get to work as distance increases.

Indirect random utility for individual i of type k when choosing tenure status t in neighborhood j is given as follows.

$$V_{i,j,k,t} = \beta_0 \left[\frac{1}{1-\nu} I_k^{1-\nu} - \frac{1}{1+\eta} \delta P_{t,j}^{\eta+1} \right] + \gamma_k D_{i,j} + \beta_{X,k} X_j + \beta_{Z,k} Z_j + \xi_{j,k,t} + \varepsilon_{i,j,k,t} \tag{1}$$

The functional form for the component of indirect utility that involves income and neighborhood price, $\left[\frac{1}{1-\nu} I_k^{1-\nu} - \frac{1}{1+\eta} \delta P_{t,j}^{\eta+1} \right]$, was selected because it allows for income

and price substitution patterns that are less restrictive than say Cobb-Douglas, but is not so complicated that estimation becomes intractable. This form of indirect utility has a Cobb-Douglas limiting distribution when the price and income elasticities, η and ν, approach the values of -1 and 1 that are implicit under the assumption of Cobb-Douglas preferences. From Roy's identity, the inferred quantity of housing demand is $H_{i,j,k,t} = \delta I_{i,k}^{\nu} P_{t,j}^{\eta}$. One important component of the estimation strategy is to apply discrete choice methods to model choices over neighborhoods. Given the error structure, the framework gives rise to a multinomial logit [McFadden (1974)] model of neighborhood choice. Direct application of maximum likelihood methods at this point runs into several problems. First, the unobserved neighborhood characteristics $\xi_{j,k,t}$ are likely correlated with the price index and racial composition. For this reason, the method presented in Berry (1994) that allows an instrumental variables approach to modeling the neighborhood specific parts of random utility is used. To this end, utility can be rewritten into the components of utility that are specific to the neighborhood choice and the utility associated with the distance between work and neighborhood.

$$V_{i,j,k,t} = \theta_{j,k,t} \; + \; \gamma_k ln(D_{i,j}) \; + \; \varepsilon_{i,j,k,t} \tag{2}$$

Conceptually, maximum likelihood can be applied in a logit framework to obtain estimates of $\theta_{j,k,t}$ and γ_k, with one of the $\theta_{j,k,t}$ being normalized to location. The estimation of $\theta_{j,k,t}$ and γ_k will be referred to hereafter as the first stage of estimation. Given the maximum likelihood estimates of $\theta_{j,k,t}$, $\hat{\theta}_{j,k,t}$, from the first stage, an instrumental variables approach [Berry (1994)] can be used to estimate the utility parameters contained in the $\widehat{\theta}_{j,k,t}$. In particular, we have the following nonlinear

regression.

$$\hat{\theta}_{j,k,t} \quad = \quad \beta_0 \left[\frac{1}{1-\nu} I_k^{1-\nu} - \frac{1}{1+\eta} \delta P_{t,j}^{\eta+1} \right] \; + \; \beta_{X,k} X_j \; + \; \beta_{Z,k} Z_j \; + \; \xi_{j,k,t} \qquad (3)$$

This aspect of the model is similar, though not identical, to the framework outlined in Berry (1994). The similarity is through the inclusion of $\theta_{j,k,t}$ which implies that the maximum likelihood estimates will perfectly fit the observed shares of each type in each community. This requirement of a perfect fit at the maximum likelihood estimates facilitates, for a given set of observed shares, a unique mapping (up to location) from any potential value for the first stage parameters to the associated MLE values for the $\theta_{j,k,t}$. The approach diverges in that maximum likelihood must be applied to jointly estimate γ_k.[1]

Given identification of the housing demand parameters (η, ν, δ) which we discuss in what follows, the coefficient on $\left[\frac{1}{1-\nu} I_k^{1-\nu} - \frac{1}{1+\eta} \delta P_{t,j}^{\eta+1}\right]$ provides an avenue for separate identification of the scale parameter β_0.

When modeling exclusively by type, there will not be variation in income across neighborhood choices and thus ν, can not be identified directly from equation 3. Further, δ and η are only separately identified from functional form. However, it is possible to use our data to identify the mean rental expenditure $REXP_j$ and the mean housing value $HVAL_j$ for each of our neighborhoods. Given that an individual's

[1]Bayer, Keohane, and Timmns (2005) also use a maximum likelihood approach to estimate neighborhood specific constants like our $\theta_{j,k,t}$ along with Cobb-Douglas utility parameters that determine preferences between housing consumption and other consumption. Using individual-level census micro data, they model the choice of location in metropolitan area as well as the income for different educational groups for metropolitan areas. Choice of metropolitan area thus implies income variation with metropolitan area. Bayer, Keohane, and Timmins' inferred income variation allows identification of the Cobb-Douglas parameters through the maximum likelihood procedure. Given our framework, there is no observed income variation across location by type because income partially defines type and thus a different housing consumption/other consumption utility parameter identification strategy is developed below.

rental housing demand is given by $\delta I_j^{\nu} P_{r,j}^{\eta}$, multiplying by $P_{r,j}$, summing across all individuals in neighborhood j, and substituting in for $P_{r,j} = \rho_{r,j}$ yields a predicted value for the mean rental expenditure in community j as a function of the housing demand parameters:

$$\widehat{REXP}_j = \sum_k \left[\frac{N_{k,r,j}}{N_{r,j}} \right] \delta I_k^{\nu} \rho_{r,j}^{\eta+1}, \tag{4}$$

where $N_{r,j}$ is the number of renter households in neighborhood j and $N_{k,r,j}$ are the number of type k households renting in community j. Similarly, for owners, the predicted mean housing value as a function of the demand parameters and capitalization rates is given by:

$$\widehat{HVAL}_j = \sum_k \left[\frac{N_{k,o,j}}{N_{o,j}} \right] \delta I_k^{\nu} A_{jk}^{\eta} \rho_{o,j}^{\eta+1}, \tag{5}$$

where $N_{o,j}$ is the number of owner households in neighborhood j and $N_{k,o,j}$ are the number of type k households owning in community j.

Taking the log of equations 4 and 5 and assuming an additive mean zero error term provides the basis for two non-linear least squares estimating equations:

$$\ln(REXP_j) = \ln(\delta) + (\eta+1)\ln(\rho_{r,j}) + \ln \sum_k \left[\frac{N_{k,r,j}}{N_{r,j}} \right] I_k^{\nu} + \varepsilon_j^r \tag{6}$$

and

$$\ln(HVAL_j) = \ln(\delta) + (\eta+1)\ln(\rho_{o,j}) + \ln \sum_k \left[\frac{N_{k,o,j}}{N_{o,j}} \right] I_k^{\nu} A_k^{\eta} + \varepsilon_j^o. \tag{7}$$

3.1 Estimating Equations

To estimate the second stage parameters in a manor that accounts for cross-equation parameter restrictions, all equations are included in a single GMM estimator. First,

assuming that $Median[\xi_{j,k,t}|k,t] = 0$, equation 3 yields $KX2$ moment conditions (two for each type – one for renters and one for owners) associated with a median regression of θ on the explanitory variables. These moment conditions are of the form:

$$\frac{1}{J}\sum_{j}\left\{1\left(\hat{\theta}_{j,k,r} \;<\; \beta_0\left[\frac{1}{1-\nu}I_k^{1-\nu} - \frac{1}{1+\eta}\delta\rho_{r,j}^{\eta+1}\right] \;+\; \beta_{X,k}X_j \;+\; \beta_{Z,k}Z_j\right) - .5\right\} \tag{8}$$

and,

$$\frac{1}{J}\sum_{j}\left\{1\left(\hat{\theta}_{j,k,o} \;<\; \beta_0\left[\frac{1}{1-\nu}I_k^{1-\nu} - \frac{1}{1+\eta}\delta A_k\rho_{o,j}^{\eta+1}\right] \;+\; \beta_{X,k}X_j \;+\; \beta_{Z,k}Z_j\right) - .5\right\} \tag{9}$$

where 1(.) is the indicator function.[2] Because it is likely that the unobserved heterogeneity in these equations, $\xi_{j,k,t}$, will be correlated with both price and racial composition, instruments for these moment conditions are X_j, $\widehat{\rho}_{t,j}$, and $\widehat{Z}_j$. Where $\widehat{\rho}_{t,j}$ and $\widehat{Z}_j$ are predicted values for price and racial composition that are constructed, as described below in the implementation section, to overcome these endogeneity problems.

Next, under the assumption of mean zeros errors, equations 6 and 7 provide two additional moment conditions that serve to identify the demand parameters[3] – one

[2]Median regression is used to account for the fact that in our data, for a given type of household, there are potential choices that never occur. To account for this fact, we omit these choices from the first stage estimation. Because of the IIA characteristic of the logit model, this does not bias the estimates on the observed choices. However, we must account for the fact that these choices had such negative values for the given type that we don't observe their being chosen. In an OLS framework this is problematic. However, in the median regression, we can assume an arbitrarily low value for the theta associated with these unobserved choices – thus incorporating in our estimation the fact that these sites were considered extremely unattractive without biasing the estimates through the choice of an arbitrarily low value of theta. Additionally, the indicator function is "smoothed" using a normal CDF.

[3]Note that the moment conditions of equations 8 and 9 may also help to identify the price elasticity parameter η and the capitalization rates A_{jk}.

for renters and one for owners:

$$\frac{1}{J}\sum_{j}\left\{\ln(REXP_j) - \ln(\delta) - (\eta+1)\ln(\rho_{r,j}) - \ln\sum_{k}\left[\frac{N_{k,r,j}}{N_{r,j}}\right] I_k^{\nu}\right\} \tag{10}$$

and,

$$\frac{1}{J}\sum_{j}\left\{\ln(HVAL_j) - \ln(\delta) - (\eta+1)\ln(\rho_{o,j}) - \ln\sum_{k}\left[\frac{N_{k,o,j}}{N_{o,j}}\right] I_k^{\nu} A_{jk}^{\eta}\right\} \tag{11}$$

Moment conditions 10 and 11 do not suffer from the endogeneity problems that are inherent in the first set of moment conditions. Thus, as instruments for equation 10 we use the derivative of $\widehat{REXP}_j$ w.r.t. ν, η, and δ. Similarly, for equation 11 we use the derivatives of $\widehat{HVAL}_j$ w.r.t. ν, η, and the appreciation rates A_k. Note that because δ is not separately identified from the appreciation rates in this final moment condition, we do not include the derivative w.r.t. δ as an instrument.

3.2 The Choice to Rent: Interpreting Estimated Capitalization Rates

A main focus of this work is to incorporate into the model the choice to own or rent in a theoretically consistent manner. Our approach models this choice through the estimation of capitalization rates A_{jk} that vary by type and community. Differences across types in these capitalization rates capture the different annualized costs of borrowing to purchase a house of a given value. As suggested by Linneman (1978) and others, one would expect these costs to vary systematically across individuals of different socio-economics status due to differential access to credit markets, wealth and permanent income differences, and variation in tax advantages (marginal value of mortgage reductions). As discussed above, these appreciation rates are identified

using individual expenditure data.

It is however also possible that systematic differences in 'tastes' for owning vs. renting, separate of differential pecuniary costs for amortizing the sales price, could exist – due perhaps to issues of expected mobility or the desirability of owner occupied housing as an asset. To capture these differences in 'tastes' across types, a tenure indicator variable is included in the second stage decomposition of $\widehat{\theta}_{jkt}$. The parameter on this indicator variable is identified off of the choices individuals make at the extensive margin to own or rent.

Finally, as is discussed below in the implementation section, we allow for the effective appreciation rates to vary across communities as a function of the observed appreciation rate of owner occupied housing in each community. Here we are using actual concurrent appreciation as a proxy for expected appreciation. Poterba (1992) and others argue that expected appreciation is a key component of capitalization rates.

4 Data

This section of the paper discusses the construction of the data used in estimating the model. The empirical exercise requires three different types of data. First, the model requires data on the distribution of household income/race types across neighborhoods along with their job locations. Second, communities must be defined and separate price indices estimated for owner and rental housing in each of these communities. Finally, neighborhood attribute data for each of these communities must be constructed. Each of these three data tasks are discussed in turn.

4.1 Individual Choice Set Data

As described previously, our structural model incorporates heterogeneity in tastes by income-race groups. We use 6 income groups (0-15k, 15-35k, 35-50k, 50-75k, 75-100k, 100k+)[4] and 5 racial groups (non-Hispanic Whites, Hispanics, non-Hispanic Blacks, non-Hispanic Asians, and non-Hispanic other). Our model also involves distance to job location as an exogenous attribute. To estimate the model, we thus require the joint distribution of income, race, and job location in each of our communities (defined by location and housing tenure).

The joint demographic distribution is available from the Census Public Use Micro Sample (PUMS) data, but not at the spatial scale of our communities. It is also available from the restricted census data centers, but these data were not available to us. Accordingly, we imputed these data from the PUMS data, standard census files, and Census Transportation Planning Package (CTPP). From these standard files, we know at the Census block level the number of households in each racial group by housing tenure. At the tract level, we also know the number of households in each racial group by income. Furthermore, from the CTPP, we know, in each tract, the number of households working in each job location, by income group. Job locations are aggregated up to 24 Place-Of-Work Public Use Microdata Areas (PUMAs) in Southern California, plus Northern California, for a total of 25 discrete locations.[5] We can thus aggregate block- and tract- level data to determine marginal distributions for our communities.

Still missing is the full joint distribution of race, income, job location, and housing

[4] In the results tables below, we refer to these as Inc1 - Inc6 respectively.

[5] To obtain one job location per household, we use the job location of the designated "householder," unless that member is not working, in which case we take the location of the next-closest relative in the household (generally a spouse). If neither member is working, we assign the household to the category "no job," with a distance of zero to all locations. In future models, we will introduce additional heterogeneity in tastes for non-working households.

tenure. We impute this distribution using a constrained minimum distance estimator. Essentially, the estimator approximates the share of households, conditional on race and tenure, who fall in each income group and who work at each job location. The approximation is designed to match as closely as possible the corresponding conditional shares in the PUMA in which each community falls. (A weighted average of PUMAs is used for communities that fall in multiple PUMAs, with weights corresponding to the fraction of the community's race-tenure group falling in the PUMA.) The approximation is further constrained to exactly match the additional marginal distributions described above (race/income, and job/income). We thus estimate the joint distribution from a slightly wider geographic area while exactly matching three bivariate marginal distributions.[6]

These data provide two opportunities for gauging the accuracy of our imputation. First, suppose we had simply imputed the income-job-location distributions from the PUMA data, conditional on race and tenure, without using the additional constraints. We can compare the predicted marginal race-income distribution under this simpler imputation to the actual race-income distribution. Doing so, we find that the median absolute percentage error among the 8,370 imputations (5 races * 6 income groups * 279 communities) is 2.9 percent, and the 90th percentile is 10 percent. Thus, we can come within a 10 percent error in 90 percent of the cases, even without using all the census data. Second, we can compare our job location / race marginal distribution from our final prediction to a limited distribution of job location and White/minority status also available from the CTPP.[7] We find that the median absolute percentage

[6]One additional complicating factor is that Census racial groups are non-Hispanic White, Hispanics, all Blacks, all Asians, and all others. (I.e., in contrast to our groups, the census minority groups combine Hispanics and non-Hispanic ethnicities.) The total number of non-White Hispanics is also given. Accordingly, we similarly impute the share of each minority (Black, Asian, other) who are Hispanic in each community, minimizing the distance to the share in the LA metro area, and exactly matching the total number of non-White Hispanics in each community.

[7]We had originally considered using this additional data in our imputation, but found that it

error among the 7,254 imputations (26 job locations * 279 communities) is under 0.1% for both Whites and minorities and that the 90th percentile of errors is 0.7% and 0.6% respectively. Thus, our imputation appears to be a reasonable way to use publicly available data and obtain finer spatial resolution than at the PUMA level.

4.2 Community Definition and Price Index Construction

The study area is the Los Angeles metropolitan area. It includes portions of Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties. Neighborhoods are defined to approximate public high school attendance zones for the 1999-2000 academic year. The set of schools considered results in 279 neighborhoods. The constructed neighborhoods are built up from US Census blocks. Using GIS, each block centroid is attached to a high school based on proximity conditional on the school and block centroid being situated within the same school district.[8]

Neighborhood price indices were calculated by tenure for each neighborhood. For owner-occupied housing units, property sales data were purchased from Fidelity National Data Services (FNDS) through SiteXdata.com. This data service provides household level data including date of last sale and corresponding sales price. Household specific characteristics include type of dwelling[9], square footage, number of bedrooms, number of bathrooms, year built, and lot size. In addition, the census block identifier is attached to each housing observation, allowing it to be placed within our communities and assigned a distance to the coast.

Data was filtered by date of sale and availability of household characteristics.

was impossible to match both this distribution and the Census race-income distribution because of small rounding differences and different census weighting schemes across data sets. However, the differences are small enough that the above ex post comparison is a reasonable check on the results.

[8]This approach is also taken by Bayer, Ferreira, and McMillan (2003), who find it yields results similar to alternative approaches.

[9]Type of dwelling refers to single family dwelling as opposed to multi-family dwelling/condominium.

Table 1: Housing Data Criteria

Variable	Range	
	Min	Max
Sale Type	Full Transfer Only	
Recording Date	Jan. 1, 2000	Dec. 31, 2000
Building Size (sqft.)	20	100000
# of Bedrooms	1	25
# of Bathrooms	0	25
# of Units	0	500
Year Built	1900	2000
Lot Size (acre)†	0.01	150
Sales Prices ($)	1	99999999

†Lot size was set to zero for all multi-family dwellings and/or condos.

All data meeting the criteria described in Table 1 were collected by county for the five county area of study.[10] In total, the owner-occupied data set consists of 90,478 observations for properties sold in 2000.

Data on rental housing units from the 2000 US Census was obtained from Integrated Public Use Microdata Series (IPUMS).[11] IPUMS provides access to Census microdata. The data represents a 5% sample of the long-form 2000 Census. In terms of geographic resolution, each observation can be identified down to the PUMA level. Each observation was imputed to communities within the PUMA based on block-group-level data. In particular, building up from block-group data, we know the share of all houses within each PUMA, by bedroom and rent, that fall into each of our communities. Individual rental units in each PUMA are assumed to be in all overlapping communities, with a weight given by these shares.

For the area of study, 109,266 rental observations were available. Housing characteristics are controlled for in the development of the price indices. Therefore, housing

[10]Of the collected data, approximately 10% of the observations either fell out of our study area or had inaccurate Census Block IDs. These observations were dropped.

[11]Data is available at http://www.ipums.org

unit variables were generated such that the two data sets were consistent. Table 2 provides summary statistics.

Rental and owner occupied housing price indices were estimated jointly to guarantee identical associated quantity indices, but no restrictions were placed on the relative levels of the two price indices. Summary statistics for housing expenditures are presented in the bottom of table 2, and the statistics for estimated housing price indices are presented in the bottom of table 3.

Finally, because of its correlation with important neighborhood characteristics, in particular ozone pollution, and because it varies on a much smaller spatial scale than is captured by our community definitions, distance to coast was included as a housing attribute in the estimation of our price indices.

Housing appreciation rates were also approximated using property value data from FNDS. To supplement the observations for transactions recorded in 2000, additional observations were purchased for transactions recorded during 1999 and 2001. In total 123,280 observations were utilized - 17,073, 90,478, and 15,729 for the years 1999, 2000, and 2001 respectively. A continuous time variable was generated in order to calculate neighborhood specific appreciation rates.

4.3 Neighborhood Attributes

Data on a number of neighborhood attributes were collected as controls, including education quality, crime, ozone pollution, and the distance to downtown Las Angeles.

Neighborhood education quality is proxied using the Standard Testing and Reporting (STAR) results provided by the California Department of Education.[12] High school test scores were collected for the 1999-2000 academic year. The data contains results for a variety of subjects and reporting methods. For the purpose of this

[12]Data available at http://star.cde.ca.gov.

Table 2: Housing Data Descriptive Statistics

Variable	Description	Mean		
		All	Owner-occupied	Rental
		(199744 obs.)	(90478 obs.)	(109266 obs.)
bdrms_1	1 bedroom	0.209	0.026	0.361
bdrms_2	2 bedrooms	0.282	0.240	0.317
bdrms_3	3 bedrooms	0.257	0.430	0.114
bdrms_4	4 bedrooms	0.126	0.243	0.028
bdrms_5	5+ bedrooms	0.030	0.060	0.004
age_2	1<age<=5	0.023	0.022	0.023
age_3	5<age<= 10	0.049	0.044	0.053
age_4	10<age<= 20	0.160	0.164	0.156
age_5	20<age<= 30	0.192	0.173	0.209
age_6	30<age<= 40	0.177	0.145	0.203
age_7	40<age<= 50	0.181	0.205	0.162
age_8	50<age<= 60	0.090	0.091	0.090
age_9	age>60	0.105	0.114	0.098
acres_1[†]	Less than 1 acre	0.518	0.808	0.277
acres_2	1.0 to 9.9 acres	0.016	0.012	0.019
acres_3	10 acres or more	0.001	0.000	0.002
own_flag	1 if owner-occupied	0.453	1.000	0.000
SFD	1 if detached SFD	0.535	0.820	0.299
price	price/annual rent[††]		279,496	8,681
S.D.			232,806	4,210
Min			10,000	48
Max			7,500,000	25,200

[†]Lot size was set to zero for all multi-family dwellings and/or condos. [††]Actual transaction price for owner occupied housing. Reported monthly rent X 12 for rental housing.

research, we utilized 11th grade mean scaled scores for math and reading. School quality is represented as the sum of these two scores. Observations were missing for 4 of the 279 schools. Values for these schools were predicted using 9th and 10th grade scores.

Crime data is available through the California Office of the Attorney General.[13] The smallest resolution for the data is at the jurisdiction level and is available by county. For a measure of crime we use the FBI Crime Index normalized by population. The FBI Crime index index is an aggregate count of crimes including counts for homicide, forcible rape, robbery, aggravated assault, burglary, motor vehicle theft, larceny-theft and arson. To assign these data to our communities, each census block is assigned the crime rate of its jurisdiction, and each community is given the population-weighted average of its blocks.

Ozone measures were developed by attaching to each census block the distance weighted average of the number of exceedances of the Federal 8-hour ozone standard at its three nearest ozone monitors.[14] For each community the population weighted average of the block level measures are then aggregated to yield the community level ozone measure. Summary statistics for the public good measures are presented in table 3.

5 Preliminary Results

For the final version of this work, our plan is to estimate the complete model for thirty separate types of individuals (6 income classes by 5 race classes) with all second stage moment conditions included in one unified GMM estimator. Here, we present very

[13] Data available at http://caag.state.ca.us/cjsc/pubs.htm.

[14] Comprehensive air quality data are available at http://www.epa.gov/air/data. In future work, we will also consider particulate pollution

Table 3: Summary Statistics for Neighborhood Attributes

Variable	Mean	Std. Dev.	Min	Max
Percent White	0.4960	0.2524	0.0081	0.9204
Percent Black	0.0698	0.1160	0.0022	0.7465
Percent Hispanic	0.3000	0.2173	0.0230	0.9373
Percent Asian	0.1101	0.1098	0.0026	0.5720
Percent Other	0.0241	0.0092	0.0018	0.0895
Ozone	8.05	10.76	0.00	45.44
Education	1,405	31	1,347	1,544
Crime	0.0498	0.1272	0.0117	1.9881
Price Indices				
Owner Occupied	157,886	72,680	63,442	588,717
Rental	7,876	1,770	3,944	13,213

preliminary estimates, with the two sets of moment conditions associated with the decomposition of θ estimated sequentially.

5.1 First Stage Estimation

First stage estimates were based on equation 2. To allow for differences in tastes between employed and unemployed households within each race-income group, public good levels where interacted with an unemployment indicator and included in the fist stage estimate. Tables 4 thru 6 summarize the correlation across types for the community fixed effects and also summarize the mean difference between the fixed effect for owning and renting within in each community - aggregated for each household type.

5.2 Second Stage Estimation

Our second stage estimates η, the price elasticity of housing demand; ν, the income elasticity of housing demand; δ, the housing demand shifter; and A_{jk}, type k's an-

Table 4: Theta Correlations

White						
	Inc 1	Inc 2	Inc 3	Inc 4	Inc 5	Inc 6
Inc 1	1					
Inc 2	0.9317	1				
Inc 3	0.8756	0.9464	1			
Inc 4	0.8076	0.9055	0.9713	1		
Inc 5	0.7201	0.8383	0.9163	0.9599	1	
Inc 6	0.6292	0.7473	0.8336	0.8953	0.9362	1
Hispanic						
	Inc 1	Inc 2	Inc 3	Inc 4	Inc 5	Inc 6
Inc 1	1					
Inc 2	0.8193	1				
Inc 3	0.6671	0.831	1			
Inc 4	0.3613	0.5861	0.7141	1		
Inc 5	0.1529	0.3413	0.5936	0.7209	1	
Inc 6	0.04	0.1856	0.4136	0.5947	0.769	1
Black						
	Inc 1	Inc 2	Inc 3	Inc 4	Inc 5	Inc 6
Inc 1	1					
Inc 2	0.8413	1				
Inc 3	0.7267	0.8311	1			
Inc 4	0.5763	0.708	0.8203	1		
Inc 5	0.4304	0.5119	0.6348	0.7874	1	
Inc 6	0.3525	0.4755	0.5682	0.7586	0.7379	1
Asian						
	Inc 1	Inc 2	Inc 3	Inc 4	Inc 5	Inc 6
Inc 1	1					
Inc 2	0.7545	1				
Inc 3	0.5589	0.7983	1			
Inc 4	0.3591	0.6323	0.7886	1		
Inc 5	0.1797	0.4247	0.6425	0.818	1	
Inc 6	0.087	0.3032	0.5325	0.7478	0.8517	1

Table 5: Theta Correlations II

Income 1					
	White	Hispanic	Black	Asian	Other
White	1				
Hispanic	0.5024	1			
Black	0.3659	0.5939	1		
Asian	0.4511	0.4146	0.3614	1	
Other	0.395	0.8696	0.6442	0.4348	1
Income 4					
	White	Hispanic	Black	Asian	Other
White	1				
Hispanic	0.4264	1			
Black	0.4533	0.3941	1		
Asian	0.6817	0.4251	0.3279	1	
Other	0.5166	0.7874	0.6514	0.389	1
Income 6					
	White	Hispanic	Black	Asian	Other
White	1				
Hispanic	0.7581	1			
Black	0.4518	0.4826	1		
Asian	0.7778	0.7037	0.3835	1	
Other	0.6616	0.8708	0.5963	0.5956	1

Table 6: Mean $\theta_{OWN} - \theta_{RENT}$

	Inc 1	Inc 2	Inc 3	Inc 4	Inc 5	Inc 6
White	-0.0146	0.4073	0.5096	0.9557	1.2788	1.8728
Hispanic	-0.9741	-0.4770	0.0156	0.6292	1.4361	1.7146
Black	-1.1175	-0.9662	-0.3681	0.2353	0.7805	1.0496
Asian	-0.8219	-0.4501	0.0685	0.7635	1.4066	1.6979
Other	-1.3637	-0.9200	-0.3055	0.4092	1.0555	1.2618

Table 7: Housing Demand Parameter Estimates

parameter	interpretation	whites	hisp	blacks	asians	others
ν	income elasticity	0.223	0.158	0.110	0.143	0.148
η	price elasticity	-0.268	-0.319	-0.340	-0.266	-0.306
δ	intercept	1.440	4.120	8.610	3.200	4.080
c0	base cap. rate	0.087				
c1	cap. adj. per $ inc	-4.45E-07				
Avg. capitalization:						
	poorest group	0.087				
	inc group 2	0.078				
	inc group 3	0.070				
	inc group 4	0.062				
	inc group 5	0.051				
	richest group	0.011				

Note: all parameter estimates significant at 1% level.

nualization rate for owning in neighborhood j. It also decomposes the first-stage estimates, $\widehat{\theta}_{jkt}$, into $\beta_{X,k}$ and $\beta_{Z,k}$, the type-specific tastes for the vector of exogenous and endogenous public goods respectively, and ξ_{jkt}, the remaining unobserved residual. In the current draft, this is split into two sub-stages. In the first sub-stage, we simulataneously recover all of the capitalization and appreciation parameters based on the two sets of housing demand moments. These results are presented in table 7

5.2.1 Instruments for decomposing θ

Both price and racial composition can be expected to be endogenous to the model and correlated with the ξ_j. We instrument for racial composition by adapting the proposal of Bayer and Timmins (2005) for congestion and agglomeration in similar models. The key insight is that the pattern of neighborhood choice for each race/income group is a function of the attributes of all communities. Consequently, the racial composition of each community is likewise a function of the attributes of all other communities.

These, in turn, can be expected to be uncorrelated with ξ_j.[15] Thus, as instruments, we can use the exogenous attributes of all communities. Following Bayer and Timmins (2005), we use the structure of the logit model to determine the instrument, which is constructed as a non-linear function of these attributes. These exogenous attributes include distance to job location and its utility parameter (estimated in the first stage) and a fitted value of theta using only the exogenous attributes (ozone, TRI emissions, county dummies, a rent/own dummy, and the log of the community size in square miles [16]). (We omit crime rates and education due to the possibility that they may be endogenous to demographic composition, though we do not otherwise instrument for them in the model.)

For each individual household in the model (characterized by job location, income group, and racial group), we estimate the share choosing to live in each community based on these exogenous attributes and the logit share equations. Once they are computed for each individual household type, these shares can be aggregated up in such a way as to compute the predicted racial composition of each community. These predicted shares are based only on the exogenous attributes. The correlations between the fitted value of racial compositions from this instrument and actual compositions is 0.65 for whites, 0.52 for Hispanics, 0.44 for Asians, 0.44 for Blacks, and 0.51 for others. F-statistics for the set of race instruments, in a regression of racial composition on all the exogenous variables, are 134 for whites, 74 for Hispanics, 24 for Asians, 75 for Blacks, and 134 for others. These test statistics indicate good instruments.

[15] A similar point was made in the context of automobile demand by Berry, Levinsohn, and Pakes (1995).

[16] As shown by Ben-Akiva and Lerman (1985), when aggregate choice alternatives in a logit model contain heterogeneity in the number of more fundamental alternatives, using the log of these fundamental alternatives as a correction term is consistent with the structure of the model. Implicitly, the model states that households choose which square mile of the LA area in which to live, with all the squares miles in one of our 558 communities having the same observable characteristics but different draws from the distribution of logit errors. The correction term picks up the fact that bigger communities have more places to live. [Randy: you might want to put this elsewhere]

Our proposal for instrumenting for price follows a similar strategy, as proposed by Bayer, McMillan, and Rueben (2005). Using the same method described above, we compute aggregate demand for housing in each community based on exogenous attributes, and compute the ratio of predicted demand to actual supply, as a measure of excess demand. In a regression of the price ratio on all our exogenous variables, these instruments have F-statistics of 5.0 for the rental housing stock and 5.3 for the owner-occupied stock. Thus, our price instruments are weaker than our race instruments. In future work, we will consider simulated prices that clear markets based on the exogenous portion of demand, or ratios of demand to land area.

5.2.2 Decomposing θ

Initial results for a small number of types are presented in table 8.

6 Conclusions

Still to come.

Table 8: Theta Decompositions

White	Constant	Crime	Education	Environment	Own Dummy	Ln Sq. Miles	K	% White
100K+	-1.0361 *	0.0204 **	0.0416 **	-0.0014	-0.2002	0.0607 **	0.0005 *	0.1397 **
	(0.6779)	(0.0084)	(0.01939)	(0.00975)	(0.2824)	(0.01445)	(0.00031)	(0.03357)
50K-100K	-0.3754 **	0.0172 **	0.0164	-0.0024	0.5618 **	0.0548 **	0.0004 **	0.0997 **
	(0.1284)	(0.00489)	(0.01334)	(0.00469)	(0.2087)	(0.00886)	(0.0001358)	(0.02073)
¡15K	0.0073	0.0324 **	0.0018	-0.0011	0.1807	0.0501 **	0.0001	0.0501 **
	(0.0182)	(0.00567)	(0.01573)	(0.00385)	(0.227)	(0.0075)	(0.0001043)	(0.0184)
Hispanic	Constant	Crime	Education	Environment	Own Dummy	Ln Sq. Miles	K	% Hispanic
100K+	1.8260	0.0620	0.3072	0.0397	0.6923	0.6452 **	-0.2713	-0.9445 *
	(1.6881)	(0.1331)	(0.3532)	(0.0503)	(0.5342)	(0.2307)	(0.2213)	(0.552)
50K-100K	0.2469 **	-0.2166 **	-0.6423 **	0.0409 *	-0.3578 *	0.3060 **	-0.0931 **	-0.0167
	(0.121)	(0.073)	(0.1614)	(0.0287)	(0.1869)	(0.0672)	(0.0449)	(0.2048)
¡15K	0.0435 *	0.4515 **	-0.1270	-0.0001	0.7884 **	0.5338 **	0.1300 **	0.5480 *
	(0.0265)	(0.1098)	(0.2611)	(0.0567)	(0.275)	(0.0944)	(0.048)	(0.3186)

References

Bajari, P., & Benkard, L. C. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, *113*(6), 1239-1276.

Bajari, P., & Kahn, M. E. (2005). Estimating housing demand with an application to explaining racial segregation in cities. *Journal of Business and Economic Statistics*, *23*(1), 20-33.

Bayer, F. F., Patrick, & McMillan, R. (2003). *A unified framework for measuring preferences for schools and neighborhoods.*

Bayer, P., Keohane, N., & Timmins, C. (2005). Migration and hedonic valuation: The case of air quality. *Working Paper*.

Bayer, P., McMillan, R., & Reuben, K. (2004). An equilibrium model of sorting in an urban housing market: A study of the causes and consequences of residential segregation. *Working Paper*.

Bayer, P., & Timmins, C. (2005). Estimating equilibrium models of sorting across locations. *Working Paper*.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, *Vol.25*(No.2), 242-262.

Berry, S. T., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *Vol.63*(No.4), 841-890.

Epple, D., Romer, T., & Sieg, H. (2001, Nov). Interjurisdictional sorting and majority rule: An empirical analysis. *Econometrica*, *Vol.69*(No.6), 1437-1465.

Epple, D., & Sieg, H. (1999). Estimating equilibrium models of local jurisdictions. *The Journal of Political Economy*, *Vol.107*(No.4), 645-681.

Gillingham, R. (1983). Measuring the cost of shelter for homeowners: Theoretical and empirical considerations. *Review of Economics and Statistics*, *65*(2), 254-265.

Linneman, P. (1978). Some empirical results on the nature of the hedonic price function for the urban housing market. *Journal of Urban Economics*, *8*, 47-68.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior (from frontiers in econometrics). New York: Academic Press.

McFadden, D. (1978). Spatial interaction theory and planning models. In P. Karlqvist et al. (Eds.), (Vol. 3, p. 75-96). Amsterdam: North-Holland Publishing Company.

Poterba, J. (1992). Taxation and housing: Old questions, new answers. *American Economic Review*, *Vol.82*(2), 237-242.

Quigley, J. (1976). Housing demand in the short run: An analysis of polychotomous choice. *Explorations in Economic Research*, *3*, 76-102.

Sieg, H., Smith, V. K., Banzhaf, H. S., & Walsh, R. (2004). Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review*, *45*(4), 1044-1077.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political*

Economy, *64*, 416-424.

Walsh, R. (2004). Endogenous open space amenities in a locational equilibrium. *University of Colorado Discussion Paper in Economics*, *04-03*.