Race, Public Goods, and Neighborhood Choice

H. Spencer Banzhaf†

Joshua Sidon ‡

Randall P. Walsh ‡*

December, 2007

Extremely Preliminary Results
Please Do Not Cite or Quote

*†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 public goods and racial composition using observed residential housing choices within the Los Angeles metropolitan area. The work is part of a larger project that seeks to investigate the impact of policies that are driven by concerns over issues of environmental justice, *i.e.* disproportionate exposure of minority groups to neighborhood environmental hazards, in a housing equilibrium framework. This larger project is particularly concerned with understanding how environmentally driven gentrification and tastes for racial composition will interact to determine the distributional impacts of policies that target environmental quality in poor or minority neighborhoods.

Consider a poor urban minority neighborhood with low public good levels. A policy which improves the level of public goods, for instance an improvement in environmental quality, may result in an influx of higher income whites that are willing and able to afford this new, improved level of public goods. Conversely, preferences for communities with low minority concentrations may forestall an influx of middle class whites, leaving the original residents living in the neighborhood and consuming these new and higher levels of public goods.

Capturing these complex interactions in a simulation framework will require an empirical model of household preferences for public goods, racial composition, and tenure status (rent vs. own) that allows for preferences to vary systematically by both race and income. This paper presents an initial attempt to estimate such a model – using only publicly available data. While the goals of the larger project are very specific, implementing the research requires crossing many methodological bridges. As a result, this paper contributes to the environmental economics and urban economics literature in several important ways. First, the paper explicitly models the decision

of whether to rent or to own in a framework that allows for systematic differences in tenure behavior based on race and income. Understanding preferences for renting versus owning is crucial for evaluating the implications of large scale policy changes since over 30% of households in the U.S. do not reside in owner occupied housing.

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 public goods *and* racial preferences will provide important insights into issues of gentrification and environmental justice.

Third, the paper develops a general methodology for estimating heterogenous preferences for neighborhood attributes 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.

Finally, the results of this analysis provide important insights into the way in which preferences for public goods and various dimensions of racial composition vary by both income and race.

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 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 public goods and racial composition 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 for differential effects of environmental policy. In our model, 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).

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) style jurisdictional sorting models. This work includes Epple and Sieg (1999), and Epple, Romer, and Sieg (2001). Using 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 improvements. This work emphasizes equilibrium analysis and for this reason imposes vertical differentiation across communities in order to characterize equilibrium conditions.

The resulting models facilitate income heterogeneity and neighborhood quality preference heterogeneity, but this heterogeneity only applies to a single commonly ranked index of public goods. Though these models have been influential, they do not readily handle multiple neighborhood-specific characteristics such as racial composition and environmental quality, nor do they easily 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 (2007) . 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 both income and race, rather than education, and location choices

are made within a single metropolitan area rather than across metropolitan areas. In contrast to the mobility constraints, captured through inclusion of birth location in the model estimated by 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 identifying 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 (I classes) x race (R races), job location $l \in L$, neighborhood choice $j \in J$, and tenure choice t. Neighborhoods are characterized by price levels for owning and renting P_t, observable neighborhood attributes X, racial composition variables Z, and unobservable attributes ξ. Our notation is summarized below.

- 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.
- l_i 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, and 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, when renting and in the case of owning is equal to the 'capitalization rate' times the sales (owner occupied) price index for community 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.

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

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 components of utility that vary across individuals. To facilitate estimation, separate models are estimated for each of 5 racial groups. To allow for heterogeneity in tastes across income groups, interactions between income and both observed neighborhood attributes and tenure status are incorporated in the model. For identification, we exclude interactions with the lowest income class.

$$V_{i,j,k,t} = \theta_{j,k,t} \; + \; \gamma_k ln(D_{i,j}) \; + \; \beta_{XI}(X * Inc) \; + \; \beta_{OwnI}(OWN * Inc) \; + \; \varepsilon_{i,j,k,t} \quad (2)$$

Conceptually, maximum likelihood can be applied in a logit framework to obtain estimates of $\theta_{j,k,t}$, γ_k, β_{XI}, and β_{OwnI} – with one of the $\theta_{j,k,t}$ being normalized to location. The estimation of these parameters 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.

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

Note that, because of the exclusion of income interactions for the lowest income groups in the first stage, the coefficients on X and the ownership dummy should be interpreted as the coefficient for the lowest income group. For other income groups, these coefficients must be added to the parameter estimate for the appropriate first stage interaction term.

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 from Berry(1994) in that maximum likelihood must be applied to jointly estimate the first stage parameters.[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

[1]Bayer, Keohane, and Timmns (2007) 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.

separate identification of the scale parameter β_0.

Our modeling approach implicitly assumes that income is independent of location choice. As a result, ν falls out of all choice comparisons and 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 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:

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

where $N_{o,j}$ is the number of owner households in neighborhood j, $N_{k,o,j}$ are the number of type k households owning in community j, and A is the capitalization rate. Note, this relationship also allows us to identify the capitalization rate.

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

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

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+) 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.[2] 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 tenure. We impute this 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 interpolate the joint distribution from a slightly wider geographic area while exactly matching three bivariate marginal distributions.[3]

[2]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.

[3]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

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.[4] We find that the median absolute percentage 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

Our 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

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.

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

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.[5]

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[6], 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. All data meeting the criteria described in Table 1 were collected by county for the five county area of study.[7] 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).[8] 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

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

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

[7]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.

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

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.

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

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.

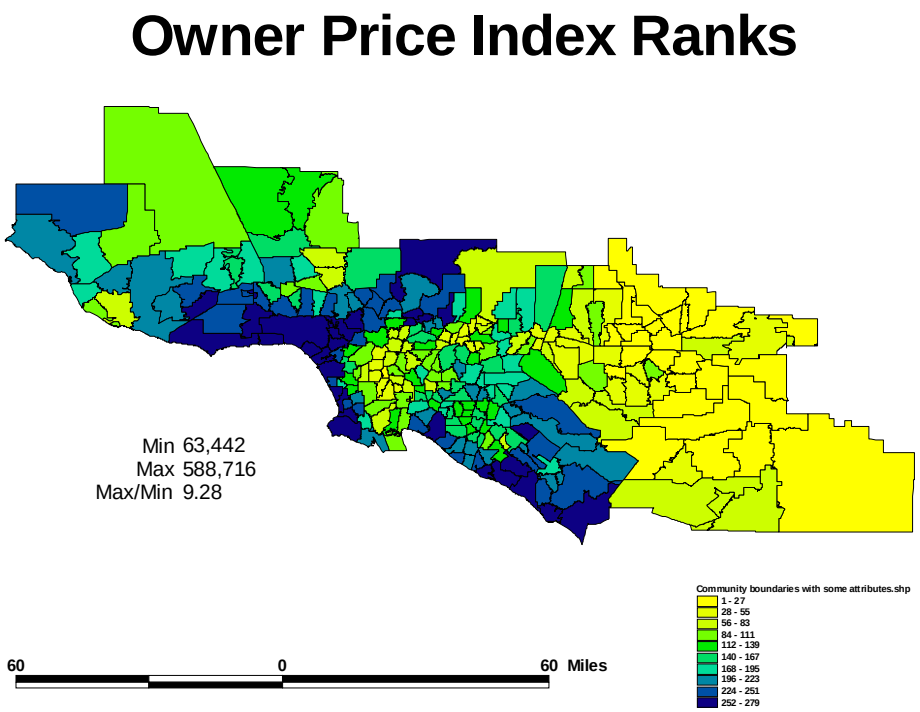
Rental and owner occupied housing price indices were estimated jointly to guarantee identical associated quantity indices, but no restrictions were placed on the relative levels of the two price indices. The spatial distribution of the price indices is presented in figure 1 and the relationship between them is presented in figure 2.

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.

Figure 1: Spatial Dispersion of Prices



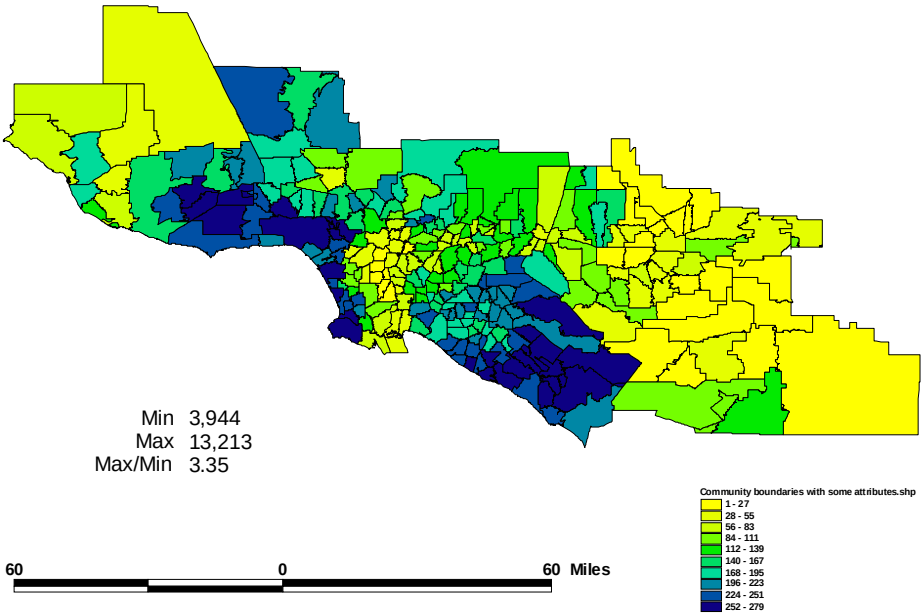
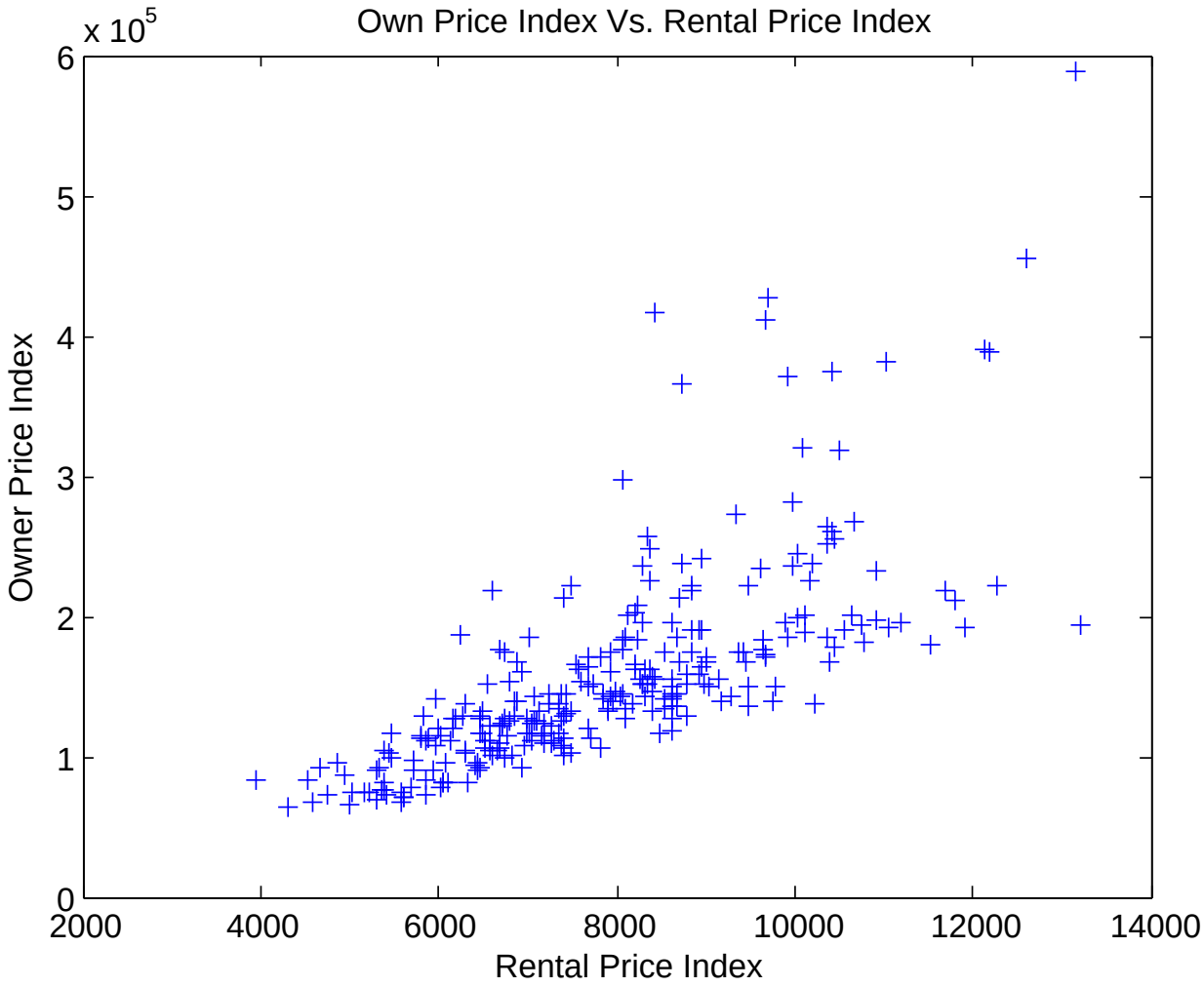


Figure 2: Price Indices:Renter Vs. Owner



Correlation = .72

4.3 Neighborhood Attributes

Data on a number of neighborhood attributes were collected as controls, including education quality, crime, and ozone pollution.

Neighborhood education quality is proxied using student teacher ratios for the 1999-2000 academic year taken from the National Center for Education Statistic's Common Core of Data.

Crime data was obtained through the California Office of the Attorney General.[9] The smallest resolution for the data is at the jurisdiction level and it is available by county. For our 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.[10] 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.

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

[10] 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

5 Results

As discussed above, estimation takes place in two stages. The first stage is a multinomial logit comprised of 558 fixed effects (one for each neighborhood/tenure choice), the parameter on distance to job location, and a set of income interactions (ozone, education, and tenure). An indicator for unemployment is also interacted with these 3 variables. A separate first stage model is estimated for each racial group. Because of the large number of fixed effects, simple maximum likelihood can not be used. Instead, estimation is implemented via concentrated maximum likelihood using the share inversion routine from Berry(1994) to solve for the vector of fixed effects that is implied by a given guess for the remaining parameters.

First we consider the community level fixed effects. Table 4 reports the correlation across racial groups for these fixed effects. The highest correlations are between blacks, hispanics and others and between whites and asians. Next, table 5 presents the coefficient estimates for the distance parameter and interaction terms. While these parameter estimates are difficult to interpret without the second stage regressions, the

Table 4: Fixed Effect Correlations

	Asian	Black	Hispanic	Other	White
Asian	1				
Black	0.5027	1			
Hispanic	0.5416	0.776	1		
Other	0.5138	0.8368	0.9778	1	
White	0.688	0.5097	0.5791	0.5154	1

pattern of the parameters are generally consistent with expectations. Higher incomes are associated with larger parameters on the public goods and stronger preferences for owning vs. renting. Note that for easier interpretation, the ozone and education variables have been normalized to a standard deviation of 1 and mean of zero.[11]

Next we consider estimation of the second stage parameters. While the fixed effect strategy in the first stage of estimation allays concerns regarding endogeneity, in stage 2, this is a serious concern. Both price and racial composition can be expected to be endogenous to the model and correlated with the unobserved community attributes ξ_j. Our instrumental variables strategies capitalizes on two characteristics of the locational equilibrium framework. First, for each race/income group, their distribution across location choices is a function of the attributes of *all* communities. Consequently, the racial composition of (and, in aggregate, the demand for) 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. Thus, as instruments, we can use the exogenous attributes of all communities. Second, the structure of the logit model implies a non-linear relationship, inherent in the model, that facilitates identification. In particular, we know that the racial composition of each community will

[11]The results presented here are quite preliminary. One shortcoming arising from the preliminary nature of the results is that while crime is included in stage 2, no interaction effects between crime and income are as yet included in stage 1.

Table 5: First Stage Parameter Estimates

	Asian	Black	Hispanic	Other	White
Distance to Job	-0.1178	-0.1175	-0.1196	-0.1122	-0.1081
	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0001)
Distance to Job (N. Cal)	-0.0045	0.0079	-0.0007	0.0003	-0.0107
	(0.0005)	(0.0006)	(0.0004)	(0.0004)	(0.0002)
Student Teacher Ratio X Inc 2	-0.0576	-0.0494	-0.0037	-0.0231	-0.0143
	(0.0046)	(0.0040)	(0.0039)	(0.0030)	(0.0025)
Student Teacher Ratio X Inc 3	-0.1041	-0.0499	-0.0356	-0.0441	-0.0310
	(0.0052)	(0.0050)	(0.0046)	(0.0036)	(0.0028)
Student Teacher Ratio X Inc 4	-0.1237	-0.0360	-0.0236	-0.0465	-0.0319
	(0.0051)	(0.0051)	(0.0048)	(0.0037)	(0.0027)
Student Teacher Ratio X Inc 5	-0.1272	-0.0565	-0.0173	-0.0382	-0.0426
	(0.0058)	(0.0063)	(0.0058)	(0.0048)	(0.0029)
Student Teacher Ratio X Inc 6	-0.1540	-0.0093	-0.0073	-0.0534	-0.0784
	(0.0054)	(0.0062)	(0.0059)	(0.0049)	(0.0027)
Ozone X Inc 2	-0.0759	0.0695	-0.0968	-0.0483	-0.0433
	(0.0093)	(0.0054)	(0.0055)	(0.0043)	(0.0026)
Ozone X Inc 3	-0.0944	0.0143	-0.1167	-0.0479	-0.0986
	(0.0108)	(0.0075)	(0.0068)	(0.0051)	(0.0031)
Ozone X Inc 4	-0.1654	-0.0710	-0.1327	-0.0909	-0.1487
	(0.0107)	(0.0080)	(0.0070)	(0.0053)	(0.0031)
Ozone X Inc 5	-0.1879	-0.0155	-0.2207	-0.1457	-0.1801
	(0.0124)	(0.0098)	(0.0093)	(0.0073)	(0.0038)
Ozone X Inc 6	-0.2486	-0.1280	-0.2828	-0.2302	-0.3440
	(0.0114)	(0.0101)	(0.0092)	(0.0076)	(0.0035)
Own X Inc 2	0.7346	0.6964	0.6541	0.6877	0.6892
	(0.0115)	(0.0112)	(0.0098)	(0.0085)	(0.0052)
Own X Inc 3	1.3337	1.4056	1.2624	1.4589	1.0885
	(0.0123)	(0.0126)	(0.0107)	(0.0090)	(0.0057)
Own X Inc 4	1.9729	2.0694	1.9632	2.1276	1.5676
	(0.0118)	(0.0126)	(0.0108)	(0.0090)	(0.0055)
Own X Inc 5	2.6956	2.6903	2.6852	2.7130	2.0833
	(0.0137)	(0.0154)	(0.0137)	(0.0111)	(0.0062)
Own X Inc 6	3.2318	3.0278	2.9541	2.8423	2.8474
	(0.0136)	(0.0159)	(0.0143)	(0.0116)	(0.0061)
Student Teacher Ratio X UnEmpl	0.0064	-0.1502	-0.0237	-0.0305	-0.0008
	(0.0032)	(0.0033)	(0.0028)	(0.0022)	(0.0016)
Ozone X UnEmpl	-0.8446	-0.6750	-0.6945	-0.6467	-0.6249
	(0.0068)	(0.0059)	(0.0048)	(0.0038)	(0.0023)
Own X UnEmpl	-0.1126	0.4797	0.1409	0.0627	0.8171
	(0.0076)	(0.0084)	(0.0064)	(0.0052)	(0.0036)
N	523321	421961	610174	951316	2466863

be determined by the predicted probabilities from the logit share equation.

Heuristically, our instruments are computed by taking the following steps. First, we regress the estimated thetas from the first stage on only the exogenous variables[12] (ozone, education, crime, county dummies, and a rent/own dummy). (We omit the endogenous racial compositions and the price term.) Together with the first-stage parameters (which are unbiased by virtue of the community fixed effect), these provide an estimate of each individual's utility function based on only exogenous data. Given these utilities, we can estimate aggregate demands for each community, as a function of housing price. Moreover, we can invert this demand function to estimate the housing price in each community that would yield the observed population profiles, conditional on the utility index for the exogenous amenities. This simulated price becomes our first instrument. Combining it with the exogenous attributes and other estimated utility parameters, we compute the simulated racial composition of each community, as well as the simulated income compositions. These simulated prices and demographics, nonlinear functions of only the exogenous data, are our instruments. Note that because of the importance of distance to job location in the first stage estimates, a large part of our identification comes from the non-uniform distribution of job locations across space.[13]

F-statistics for the price instrument, in a regression of the transformed price term on all the exogenous variables, is 12.1. The respective statistics for each racial group are 21.5 for whites, 17.2 for Hispanics, 9.3 for Asians, 2.2 for Blacks, and 22.1 for others. These test statistics indicate good instruments, with the exception of the Black racial group.

[12]Note: these fixed effects are themselves functions of only exogenous variables.

[13]Bayer and Timmins (2007) discuss this I.V. strategy for dealing with issues of congestion and agglomeration in similar models.

Table 6: Second Stage Parameter Estimates

	Asian	Black	Hispanic	Other	White
Price Term	0.1296*	0.1845*	0.1615*	0.1667*	0.1637*
	(0.0398)	(0.0499)	(0.0437)	(0.0449)	(0.0484)
Base Student Teacher Ratio	-0.1410	-0.1247	-0.1104	-0.1145	-0.1482
	(0.1139)	(0.1429)	(0.1253)	(0.1287)	(0.1387)
Base Ozone	-0.2106	-0.7209*	-0.3808	-0.5153	-0.4207
	(0.3416)	(0.4283)	(0.3757)	(0.3857)	(0.4158)
Base Own	-0.3927	-1.1716*	-0.4510	-0.8110*	-0.0277
	(0.2661)	(0.3337)	(0.2927)	(0.3005)	(0.3240)
Crime	-0.2381	-0.1743	-0.4357*	-0.4026	-0.4133
	(0.2447)	(0.3069)	(0.2692)	(0.2764)	(0.2979)
Percent Asian	.	-0.9207	49.8429*	-31.4523*	-7.8595*
	.	(8.0289)	(18.8409)	(7.6489)	3.4982
Percent Black	-13.0384*	.	46.8406*	-34.8075*	-14.5182*
	(6.4028)	.	(15.4442)	(12.7711)	(6.4465)
Percent Hispanic	-58.8952*	-65.4522*	.	-87.7034*	-71.2391*
	(17.1278)	(17.6058)	.	(26.0080)	(19.3680)
Percent Other	18.3915*	28.4323*	77.4079*	.	21.8162*
	(6.7730)	(14.1808)	(25.3331)	.	(9.1490)
Percent White	-1.1609	5.2219	55.8740*	-25.5233*	.
	(2.8734)	(6.6398)	(17.4996)	(8.4866)	.
Imputed Percent Own Race	3.4189	2.0449	-14.3728	11.2179	4.4875
Demand Parameters					
price elasticity	-0.6350*				
	(0.0342)				
income elasticity	0.4980*				
	(0.0281)				
capitalization term	0.0561*				
	(0.0010)				
demand shifter	2.2589*				
	(0.4102)				

Second stage parameter estimates are presented in table 6.[14] The parameters for all six races are estimated simultaneously using three stage least squares and instrumenting as discussed above. The first row of the table presents estimates for the scale factor β_0. This is the parameter on the term $\left[\frac{1}{1-\nu} I_k^{1-\nu} - \frac{1}{1+\eta}\delta P_{t,j}^{\eta+1}\right]$.[15] Because of the sign on price in this term, the expected sign of this coefficient is positive. The coefficient estimates are all positive and significant.

The next set of coefficients are for the public goods and ownership dummy. Recall that given the interaction terms in the first stage, these are the coefficient for the lowest income group. While few of them are significantly different from zero, they are all of the expected sign. The next set of coefficients are the coefficients on the percent of each race group in the neighborhood. Because the sum of these percentages across all race groups is by definition equal to one, one race group must be excluded. In each model, we drop the percent of own race. These coefficients are largely significant, suggesting that racial composition is important in household location choice. For comparison, we also impute the implied coefficient for a 1 percentage point increase in own race, assuming that this increase is achieved by equal reductions in the percentage of the other four race groups.

Interpreting the empirical results requires combining the first and second stage estimates to recover the coefficients for each type of individual. Aggregated coefficients are presented in Figures 3 thru 5. The Figures summarize the coefficients for ozone, student teacher ratio, and ownership by combining first and second stage estimates to identify coefficient estimates and 90% confidence intervals.

Finally, Figure 6 plots the willingness to pay for changes in ozone, student teacher ratio and percent own race. The first plot in the figure reports the marginal willingness

[14] Note: standard errors are uncorrected.

[15] The parameters imbedded in this term are estimated on equations 4 and 5. These estimates are reported in the bottom of table 6.

Figure 3: 90% Confidence Intervals - Ozone

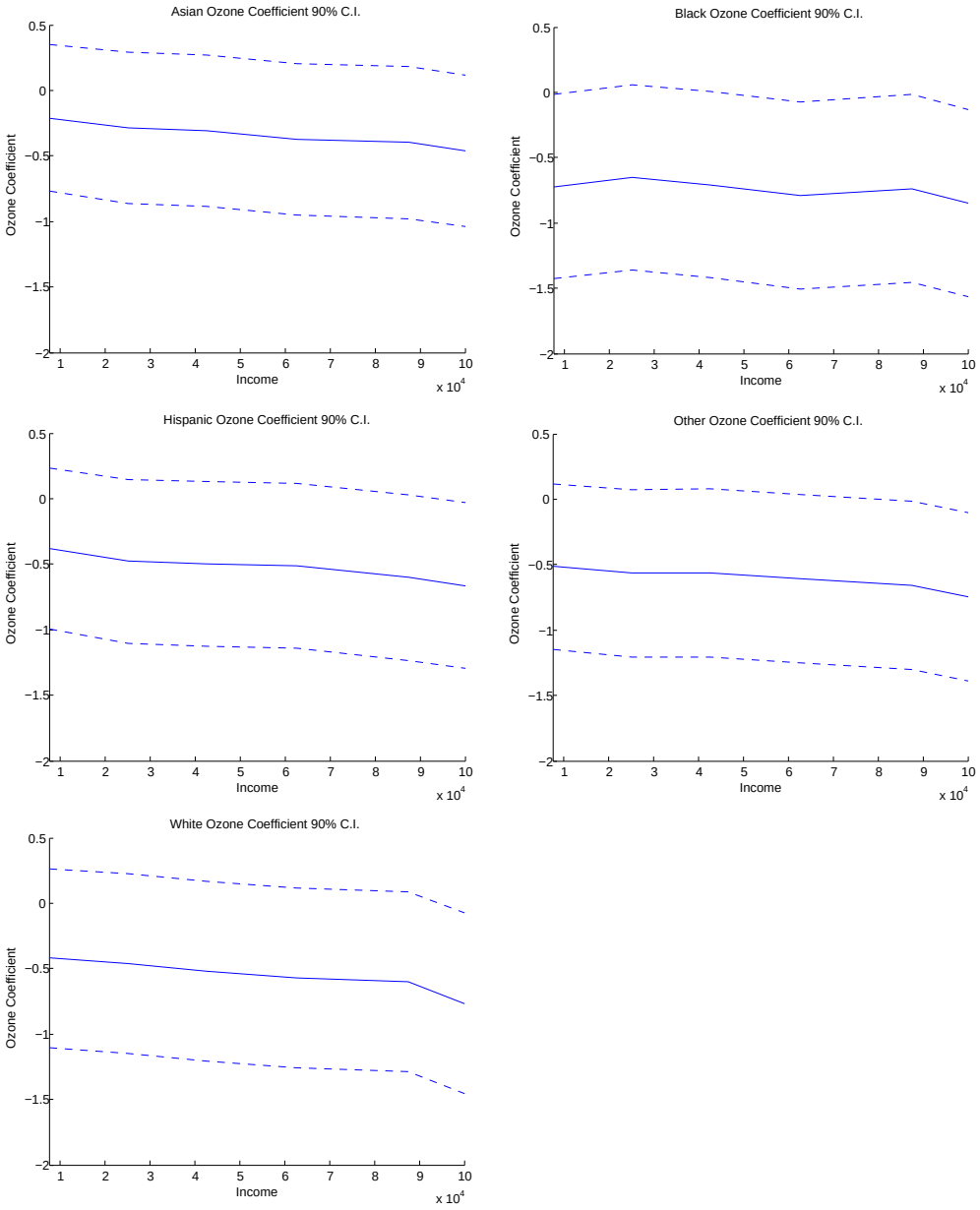


Figure 4: 90% Confidence Intervals - Student Teacher Ratio

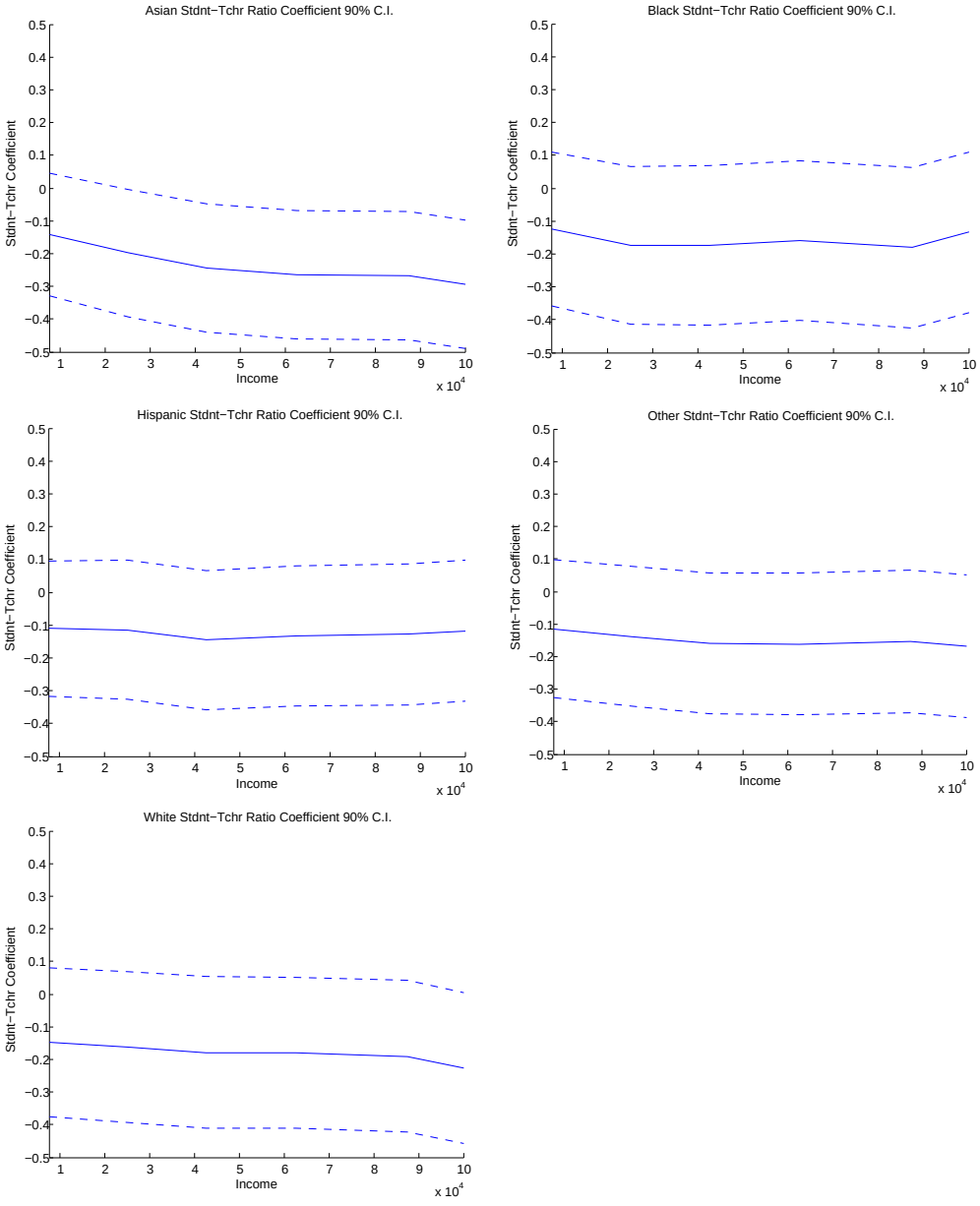


Figure 5: 90% Confidence Intervals - Ownership Indicator

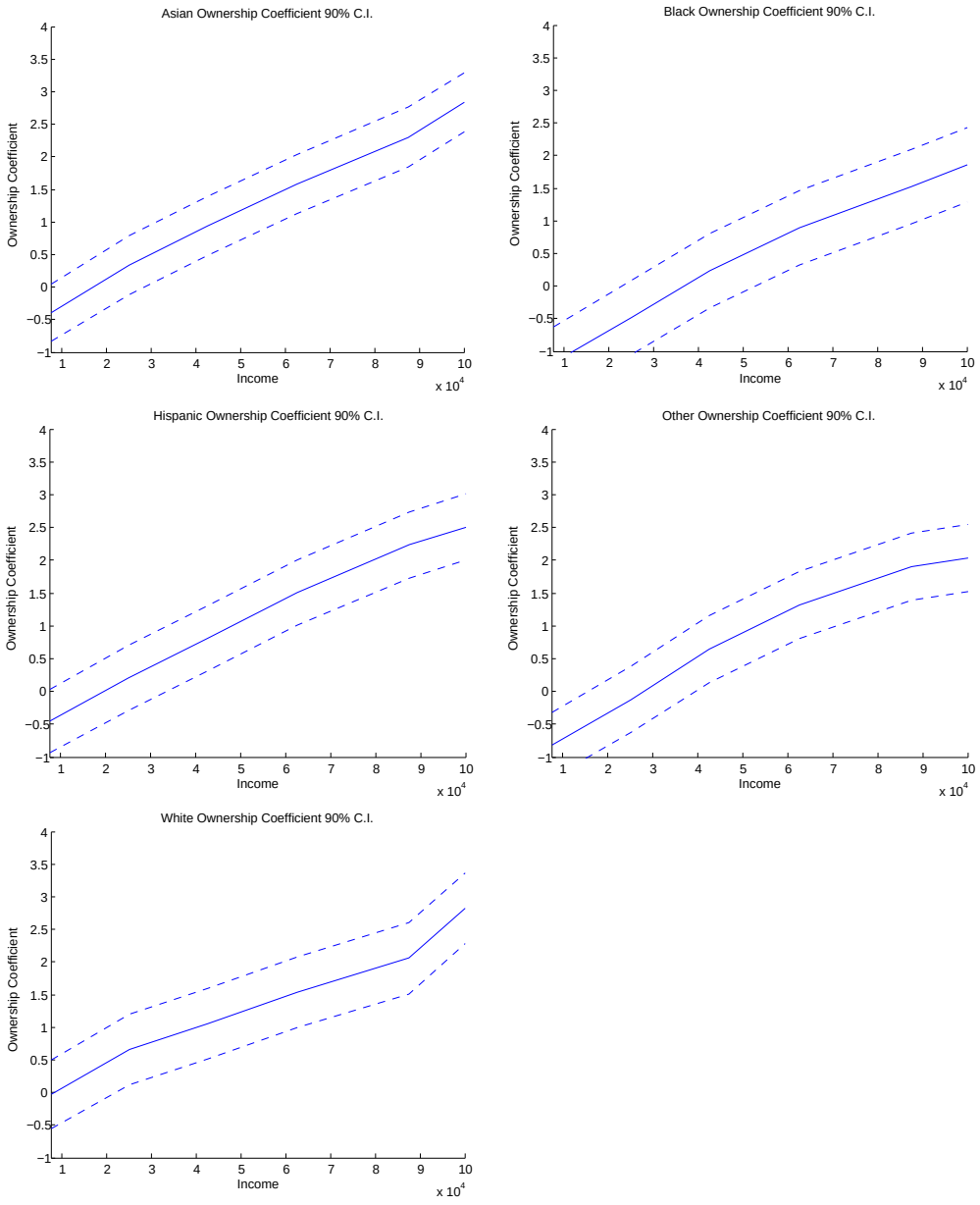
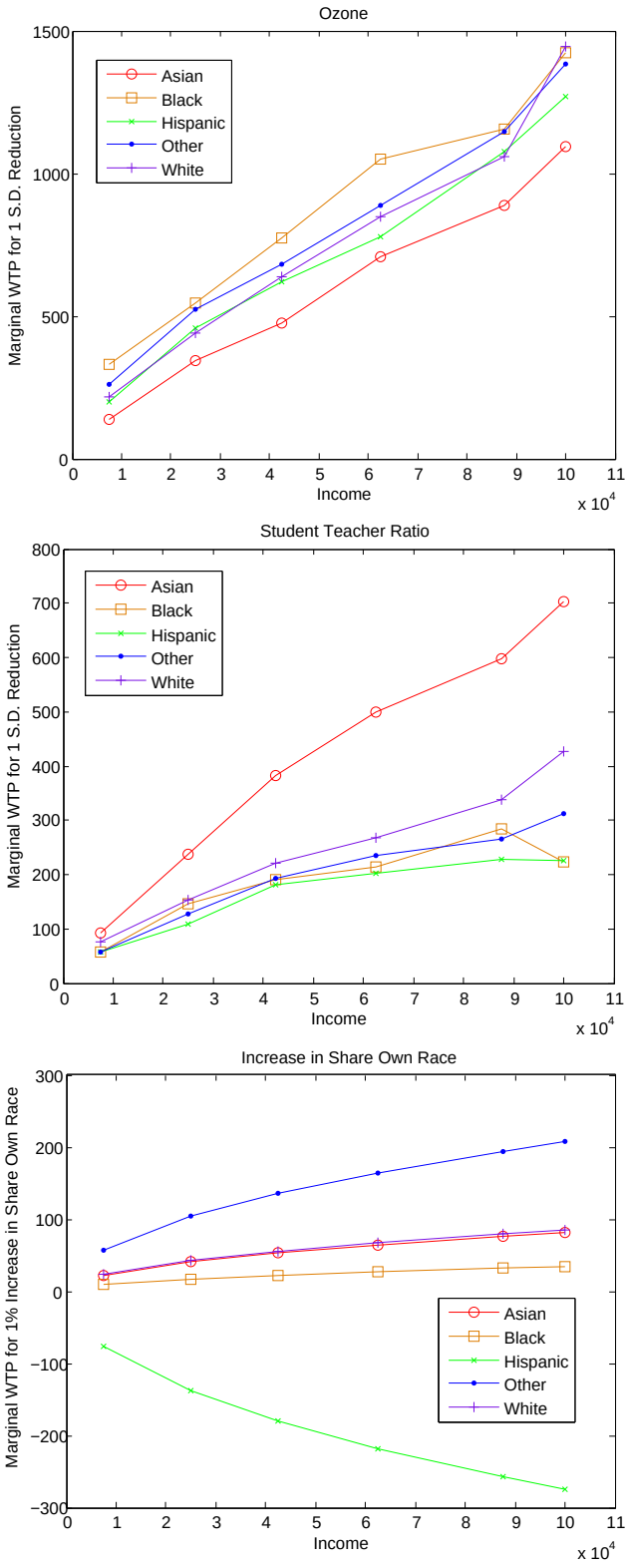


Figure 6: WTP by Income and Race



to pay for a one standard deviation in ozone levels by race and income. The results suggest very little difference across races in the income profile of willingness to pay. The second plot reports the willingness to pay for reductions in the student teacher ratio. Again, there is very little difference in the WTP profile across races. With one exception, asians appear to have markedly higher wtp levels than the other racial groups. Especially at the highest income levels, where their willingness to pay is almost twice that of other racial groups.

The final panel of Figure 6 reports the willingness to pay for a 1 percentage point increase in own race. As expected, most racial groups have a positive willingness to pay for an increase in their own racial group. The one exception is for hispanics who actually have a negative willingness to pay for an increase in their own racial groups representation in the community. Interestingly, a review of Table 6 shows that across all race groups, hispanics are viewed the least positively.

6 Conclusions

Still to come.

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*.

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.

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*.