DETAILED DESCRIPTION OF DATA USED IN:

A UNIFIED FRAMEWORK FOR MEASURING PREFERENCES FOR SCHOOLS AND NEIGHBORHOODS[+]

Patrick Bayer
Department of Economics
Yale University

Fernando Ferreira
Department of Economics
University of California, Berkeley

Robert McMillan
Department of Economics
University of Toronto

July 2003

[+] Paper available at: http://aida.econ.yale.edu/~pjb37/bfm-jpe.pdf.

1. Introduction

This document details the sources for the data and the construction of the variables used in "An Unified Approach for Measuring Preferences for Schools and Neighborhoods," by Patrick Bayer, Fernando Ferreira, and Robert McMillan.

2. Census Variables

House Prices

Because house values are self-reported, it is difficult to ascertain whether these prices represent the current market value of the property, especially if the owner purchased the house many years earlier. Fortunately, the Census also contains other information that helps us to examine this issue and correct house values accordingly. In particular, the Census asks owners to report a continuous measure of their annual property tax payment. The rules associated with Proposition 13 imply that the vast majority of property tax payments in California should represent exactly 1 percent of the transaction price of the house at the time the current owner bought the property or the value of the house in 1978. Thus, by combining information about property tax payments and the year that the owner bought the house (also provided in the Census in relatively small ranges), we are able to construct a measure of the rate of appreciation implied by each household's self-reported house value. We use this information to modify house values for those individuals who report values much closer to the original transaction price rather than current market value. In our study most households list the purchase price of their house rather than an estimated market value for their house. Thus if two identical houses were found in the census data but one was last sold in 1989 and one was last sold in 1969 we find on average the listed market price of the more recently sold house is on average 15 percent higher than the other house.

A second deficiency of the house values reported in the Census is that they are top-coded at $500,000, a top-code that is often binding in California. Again, because the property tax payment variable is continuous and not top-coded, it provides information useful in distinguishing the values of the upper tail of the value distribution. We find that top-coding was fairly predominant in the Bay Area and that higher top-codes may be useful to gain a better understanding of house prices in expensive markets like California or New York.

The exact procedure that we use to adjust self-reported house values is as follows. We first regress the log of self-reported house value on the log of the estimated transaction price (100 times the property tax payment), and a series of dummy variables that characterize the tenure of the current owner:

$$(1) \qquad \log(V_j) = \alpha_1 \log(T_j) + \alpha_2 y_j + \omega_j$$

where V_j represents the self-reported house value, T_j represents the estimated transaction price, and y_j represents a series of dummy variables for the year that the owner bought the house. If owner-estimated house values were indeed current market values and houses were identical except for owner tenure, this regression would return an estimate of 1 for α and the estimated α_2 coefficients would indicate the appreciation of house values in the Bay Area over the full period of analysis. If owners tend to underreport house values, especially when they have lived in the house for a long time, the estimated α_2 parameters will likewise underreport appreciation in the market. In this way, the estimated α_2 parameters represent a conservative estimate of appreciation. Given the estimates of equation (2), we construct a predicted house value for each house in the sample and replace the owner-reported value with this measure when this predicted measure exceeds the owner-reported value. In practice, in order to allow for different rates of appreciation in different regions of the housing market, we conduct these regressions separately for each of the 45 Census PUMA (areas with at least 100,000 people) in our sample and allow appreciation to vary with a small set of house characteristics within each PUMA. In this way, the first adjustment that we make to house prices is to adjust owner-reported values for likely under-reporting.

The adjustment to top-coded house prices uses the same approach, using the information on property taxes that are continuous and not top-coded. Using estimates of equation (2) based on a sample of houses that does not include the top-coded house values, we construct predicted house values for all top-coded houses. This allows us to assign continuous house values for top-coded measures.

Reported Rental Value

We next examined questions of reported monthly rents. While rents are presumably not subject to the same degree of misreporting as house values, it is still the case that renters who have occupied a unit for a long period of time generally receive some form of tenure discount. In some cases, this tenure discount may arise from explicit rent control, but implicit tenure discounts generally occur in rental markets even when the property is not subject to formal rent control. Thus while, this will not lead to errors in the answering of the listed census question it may lead to an inaccurate comparison of rents faced by households if they needed to move. In order to get a more accurate measure of the market rent for each rental unit, we utilize a series of locally based hedonic price regressions in order to estimate the discount associated with different durations of tenure in each of over 40 sub-regions within the Bay Area.

In order to get a better estimate of market rents for each renter-occupied unit in our sample, we regress the log of reported rent R_j on a series of dummy variables that characterize the tenure of the current renter, y_j, as well as a series of variables that characterize other features of the house and neighborhood X_j:

$$(2) \qquad \log(R_j) = \beta_1 y_j + \beta_2 X_j + \upsilon_j$$

again running these regressions separately for each of the 45 PUMAs in our sample. To the extent that the additional house and neighborhood variables included in equation (3) control for differences between the stock of rental units with long-term vs. short-term tenants, the β_1 parameters provide an estimate of the tenure discount in each PUMA.[1] In order to construct estimates of market rents for each rental unit in our sample, then, we inflate rents based on the length of time that the household has occupied the unit using the estimates of β_1 from equation (2). In this way, these three price adjustments bring the measures for rents and house values reported in the Census reasonably close to market rates.

Calculating Cost Per Unit of Housing Across Tenure Status

Finally, in order to make owner- and renter-occupied housing prices comparable in our analysis we need to calculate a current rental value for housing. Because house prices reflect the expectations about the future rents for the property they incorporate beliefs about future housing appreciation. To appropriately deflate housing values -- and especially to control for differences in expectations about appreciation in different segments of the Bay Area housing market – we regress the log of house price (whether monthly rent or house value) Π_j on an indicator for whether the housing unit is owner-occupied o_j and a series of additional controls for features of the house including the number of rooms, number of bedrooms, types of structure (single-family detached, unit in various sized buildings, etc.), and age of the housing structure as well as a series of neighborhood controls X_j:

$$(3) \qquad \log(\Pi_j) = \gamma_1 o_j + \gamma_2 X_j + \eta_j$$

We estimate these hedonic price regressions for each of 40 sub-regions (Census Public Use Microdata Areas - PUMAs) of the Bay Area housing market. These regressions return an estimate of the ratio of house values to rents for each of these sub-regions and we use these ratios to convert house values to a measure of current monthly rent.

3. External Data

We next discuss the additional variables we have added to the census data to provide a more nuanced understanding of the neighborhood characteristics that impact house prices and residential location decisions. These data sets are linked to census blocks and can be used to determine the appropriateness of the questions and sampling techniques used. This additional data includes:

School and School District Data

The Teale data center in California provided a crosswalk that matches all Census blocks in California to the corresponding public school district. We have further matched Census blocks to particular schools using a variety of procedures that takes account of the location (at the block level) of each Census block within a school district and the precise location of schools within the district using information on location from the Department of Education. Other school information in these data include:

[1] Interestingly, while we estimate tenure discounts in all PUMAs, the estimated tenure discounts are substantially greater for rental units in San Francisco and Berkeley, the two largest jurisdictions in the Bay Area that had formal

- 1992-93 CLAS dataset provides detailed information about school performance and peer group measures. The CLAS was a test administered in the early 1990s that will give us information on student performance in math, literature and writing for grades 4, 8 and 10. This dataset presents information on student characteristics and grades for students at each school overall and across different classifications of students, including by race and education of parents.
- 1991-2 CBEDS (California Board of Education data sets) datasets including information from the SIF (school information form) which includes information on the ethnic/racial and gender make-up of students, PAIF – which is a teacher based form that provides detailed information about teacher experience, education and certification backgrounds and information on the classes each teacher teaches, and (LEP census) a language census that provides information on the languages spoken by limited-English speaking students.

Procedures for Assigning School Data:

While we have an exact assignment of Census blocks to school districts, we have only been able to attain precise maps that describe the way that city blocks are assigned to schools in 1990 for Alameda County. In the absence of information about within-district school attendance areas, we employ the alternative approaches for linking each house to a school. The crudest procedure assigns average school district characteristics to every house falling in the school district. A refinement on this makes use of distance-weighted averages. For a house in a given Census block, we calculate the distance between that Census block and each school in the school district. We have detailed information characterizing each school and construct weighted averages of each school characteristic, weighting by the reciprocal of the distance-squared as well as enrollment.

As a third approach we simply assign each house to the closest school within the appropriate school district. Our preferred approach (which we use for the results reported in the paper) refines this closest-school assignment by using information about individual children living in each Census block - their age and whether they are enrolled in public school. In particular, we modify the closest-school assignment technique by attempting to match the observed fourth grade enrollment for every school in every school district in the Bay Area. Adjusting for the sampling implicit in the long form of the Census, the 'true' assignment of houses to schools must give rise to the overall fourth grade enrollments observed in the data.

These aggregate numbers provide the basis for the following intuitive procedure: we begin by calculating the five closest schools to each Census block. As an initial assignment, each Census block and all the fourth graders in it are assigned to the closest school. We then calculate the total predicted enrollment in each school, and compare this with the actual enrollment. If a school has excess demand, we reassign Census blocks out of its catchment area, while if a school has excess supply, we expand the school's catchment area to include more districts.

To carry out this adjustment, we rank schools on the basis of the (absolute value of) their prediction error, dealing with the schools that have the greatest excess demand/supply first. If the school has excess demand, we reassign the Census

rent control in 1990.

block that has the closest second school (recalling that we record the five closest schools to each Census block, in order), as long as that second school has excess supply. If a school has excess supply, we reassign to it the closest school district currently assigned to a school with excess demand. We make gradual adjustments, reassigning one Census block from each school in disequilibrium each iteration. This gradual adjustment of assignments of Census blocks to schools continues until we have 'market clearing' (within a certain tolerance) for each school. Our actual algorithm converges quickly in practice, and produces plausible adjustments to the initial, closest-school assignment.

Land use

Information on land use/land cover digital data is collected by USGS and converted to ARC/INFO by the EPA available at: http://www.epa.gov/ost/basins/ for 1988. We have calculated for each Census block, the percentage of land in a ¼, ½,1, 2, 3, 4 and 5-mile radii that is used for commercial, residential, industrial, forest (including parks), water (lakes, beaches, reservoirs), urban (mixed urban or built up), transportation (roads, railroad tracks, utilities) and other uses.

Crime data

Information on crime was drawn from the rankings of zipcodes on a scale of 1-10 on the risk of violent crime (homicide, rape or robbery). A score of 5 is the average risk of violent crime and a score of 1 indicates a risk 1/5 the national average and a 10 is 10 or more times the national average. These ratings are provided by CAP index and were downloaded from APBNews.com.

Geography and Topography

The Teale data center in California provided information on the elevation, latitude and longitude of each Census block.