Online Appendix
"The Housing Market(s) of San Diego"

Tim Landvoigt, Monika Piazzesi & Martin Schneider

January 8, 2015

A   San Diego County Transactions Data

In this appendix we describe our selection of sales and repeat sales. We begin by describing our sample of sales which not only forms the basis for selecting repeat sales but is also used to illustrate the shift in distributions in Section 2. Our goal is to compile a dataset of households' market purchases of single-family dwellings. We start from a record of all deeds in San Diego County, 1999-2008 and then screen out deeds according to three criteria.

First, we look at qualitative information in the deed record on what the deed is used for. We drop deed types that are not typically used in arms length transfers of homes to households in California. In particular, we keep only grant deeds, condo deeds, corporate deeds and individual deeds. The most important types eliminated are intrafamily deeds and deeds used in foreclosures. Even for the types of deeds we keep, the deed record sometimes indicates that the transaction is not "arms length" or that the sale is only for a share of a house – we drop those cases as well.

Second, we drop some deeds based on characteristics of the house or the buyer. We use only deeds for which a geocode allows us to precisely identify latitude and longitude. We eliminate deeds that transfer multiple parcels (as identified by APN number.) Information about property use allow us to eliminate second homes and trailers. To further zero in on household buyers, we eliminate deeds where the buyer is not a couple or a single person (thus dropping transaction where the buyer is a corporation, a trust or the beneficiary of a trust.)

Third, we drop some deeds based on the recorded price or transaction dates. We drop deeds with prices below \$15,000 or with loan-to-value ratios (first plus second mortgage) above 120%. We also consolidate deeds that have the same sales price for the same contract date. We drop deeds that have the same contract date but different prices.

Our repeat sales sample is used to estimate our statistical model of price changes in Section 2.2. A repeat sale is a pair of consecutive sales of the same property within the above sales sample. Since we are interested in long term price changes, we want to avoid undue influence of house flipping on our estimates. We thus drop all pairs of sales that are less than 180 days apart. To guard against outliers, we drop repeat sales with annualized capital gains or losses above 50%.

B. Robustness of facts on capital gains

Table B.1 reports additional results for the repeat sales model that incorporate zip code and census tract level information. Regression (i) reports the basic regression of capital gains on the own initial 2000 price p_t^i in logs from Figure 1. The regression has a slope coefficient of -0.060 with a standard error 0.0014 and an R^2 of 57.1%. Regression (ii) adds the initial zip code median (again, in logs) p_t^{zip} as regressor. The point estimate of the coefficient on the initial own price is basically unchanged (-0.057 versus -0.060); the difference is not statistically significant. The estimated coefficient on the zip-code median is statistically significant, but -0.011 is economically small. The added explanatory power of the zip code median is tiny, the R^2 goes from 57.1% to 57.5%. The regression (iii) on the zip-code median alone (iii) gives an R^2 of 20.6%. Regressions (iv) and (v) are analogous to (ii) and (iii), but they use census tract rather than zipcode as the geographical area. The results are quite similar. Regression (v) uses only the census-tract medians with an R^2 of 28.5%.

		const.	$\log p^i_{2000}$	$\log p^{zip}_{2000}$	$\log p^{census}_{2000}$	R^2
(i)	$\log p^i_{2005} - \log p^i_{2000}$	0.899 (0.018)	−0.060 (0.002)			0.571
(ii)	$\log p^i_{2005} - \log p^i_{2000}$	1.000 (0.035)	−0.057 (0.002)	−0.011 (0.003)		0.575
(iii)	$\log p^i_{2005} - \log p^i_{2000}$	1.016 (0.048)		−0.069 (0.004)		0.206
(iv)	$\log p^i_{2005} - \log p^i_{2000}$	0.879 (0.024)	−0.062 (0.002)		0.004 (0.003)	0.572
(v)	$\log p^i_{2005} - \log p^i_{2000}$	0.837 (0.031)			−0.056 (0.003)	0.285
(vi)	$\log p^{zip}_{2005} - \log p^{zip}_{2000}$	1.014 (0.068)		−0.070 (0.006)		0.672
(vii)	$\log p^{census}_{2005} - \log p^{census}_{2000}$	1.038 (0.034)			−0.071 (0.003)	0.606

Note: This table reports results from regressions of the capital gain from 2000 to 2005 in the price series indicated on the left-hand side on the regressors indicated on the headers of the columns. These cross sectional regressions involve individual house prices p^i_t from houses that were repeat sales in the two years 2000 and 2005, zip code medians p^{zip}_t, and census tract medians p^{census}_t in San Diego County.

Regression (vi) runs the capital gains in the zip-code medians on the initial zip-code medians. The results are comparable to row (i) with a slope coefficient of −0.070, a somewhat higher standard error of 0.0055, and a higher R^2 of 67.2%. Regression (vii) does the same exercise for census tracts, again with similar slope as row (i).

Figure 11 plots the repeat sales observations from Figure 1 with the 2005 log price on the y-axis. The black line is the predicted value from a linear regression of 2005 log prices on 2000 log prices. The green line is the predicted

value from a nonparametric regression, using a Nadaraya-Watson estimator with a Gaussian kernel and a bandwidth of 0.15. The nonparametric regression line is strictly increasing in the initial 2000 price. This monotonicity property implies that the relative ranking of houses by quality according to the nonparametric regression is the same as the relative ranking according to the linear regression.
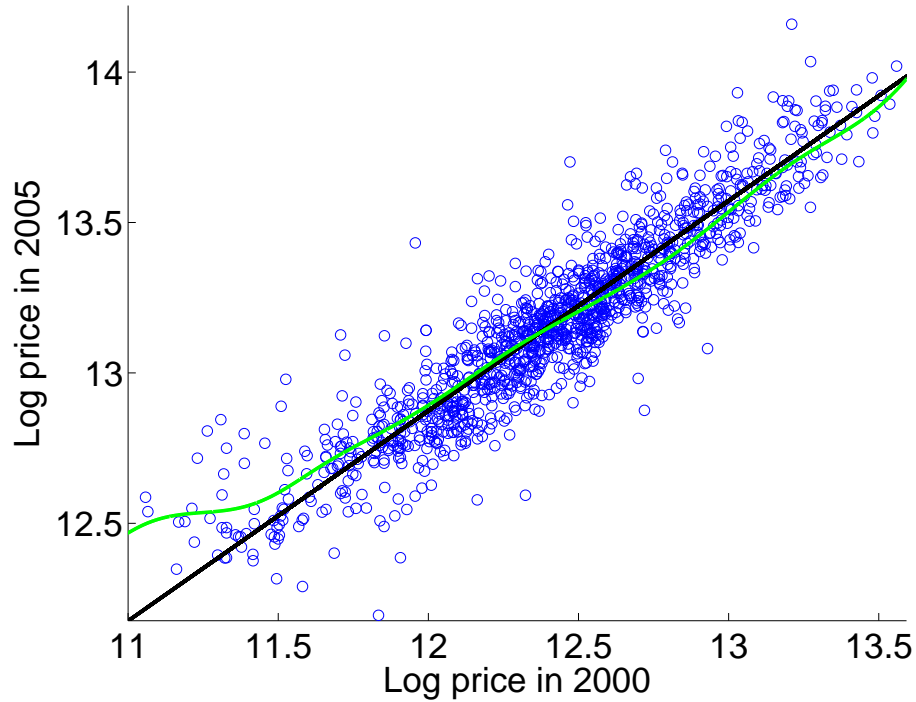

Figure 11: Repeat Sales in San Diego County, CA, during the years 2000-2005.

The nonparametric regression line is close to linear for a large range of house values, with the largest deviation at the low end. This deviation does not matter for our approach, because we use the pricing model only to derive an ordinal index. The absolute amount of service flow due to a house of a certain ordinal quality is backed out using the structural model. Section 5.1 uses 2000 house prices to back out a service flow function for that year and assumes a constant rate between 2000 and 2005 to derive the 2005 service flow function. Section E uses 2000 house prices and 2005 house prices to back out service flow functions for the two years, respectively.

4

C. Details on quantitative implementation

This appendix provides details on the calculations of home improvements, house quality and wealth reported in the text.

Improvements

The 2002 American Housing Survey contains data on home improvements in San Diego County. Table C.1 shows the means and medians of annual improvement expenses in San Diego as a percent of house values. We find that San Diego homeowners spend an amount equal to roughly 1% of their house value on improvements each year. The mean percentage spent on improvements is 2% for homes in the lowest bin, worth less than $50,000. However, this higher mean is estimated imprecisely. A test that the mean improvement percentage in the lowest bin and homes in the next bin (worth between $50,000 and $100,000) are identical cannot be rejected at the 10% level. A joint test whether mean improvements across all bins are equal can also not be rejected. We also test whether the data are drawn from populations that have the same median and cannot reject.

TABLE C.1: HOME IMPROVEMENTS IN SAN DIEGO

House Value (in thousands)

	<50	50-100	100-150	150-200	200-300	300-500	500-1,000	>1,000
Improvements (in percent)								
mean	2.11	1.05	0.73	0.75	0.95	0.96	0.8	1.1
	(0.54)	(0.32)	(0.11)	(0.09)	(0.13)	(0.12)	(0.13)	(0.30)
median	0	0.07	0.08	0.10	0.14	0.19	0.13	0.13

Note: This table contains the estimated means of home improvement measured as percent of house value. These statistics are computed for observations within the house price bins indicated on the top of the table. The data are the San Diego County observations of the 2002 American Housing Survey on 'rac' which measures the cost of replacements/additions to the unit.

Census house values

The Census data does not contain actual prices but rather price ranges, including a top range for houses worth more than one million dollars. The topcoded range contains 9.6% of houses in 2005 and 1.8% of houses in 2000. To obtain cross sectional distributions of houses sold in a given year, we fit splines through the bounds of the Census house binds. Let p^c a vector that contains those bounds, as well as a lower bound of zero and an upper bound. We can obtain a continuous distribution for every upper bound by fitting a shape-preserving cubic spline through $(p^c, G_0(p^c))$. We choose the upper bound such that the median house value in the topcoded range equals the median in that range in our transaction data. To prepare the imputation of wealth (described below) we set a household's housing wealth to the midpoint of its bin, and we use the median of the topcoded range for the top bin.

Imputation of wealth

For age and income, we use age of the household head and income reported in the 2000 Census (for $t = 2000$) and 2005 ACS (for $t = 2005$). We are thus given age and income, as well as a survey weight, for every survey household. However, Census data do not contain wealth. We construct a conditional distribution of wealth using data from the Survey of Consumer Finances (SCF). We use the 1998 and 2004 SCF to build the conditional distributions for 2000 and 2005, respectively.

We use a chained equations approach to perform imputations. The estimation is in two steps. In the first step, we use SCF data to run regressions of log net worth on log housing wealth, a dummy for whether the household has a mortgage and if yes, the log mortgage value, and log income for each age decade separately. In the second step, we use a regression switching approach described in Schenker and Taylor (1996), and implemented in the Stata commands 'mi' or 'ice'. The procedure draws simulated regression coefficients from the posterior distribution of the coefficients estimated using the SCF data. Using these simulated coefficients and the observed covariates in the

census data, a predicted value for log networth is calculated for each census household. The imputed log networth for each census household is then randomly drawn from a set of SCF households whose actual log networth is close to the predicted log networth of the missing observation.

By repeating this second step multiple times (i.e. drawing multiple sets of simulated coefficients), we can generate multiple imputed observations for each original census observation. We choose to create three different imputations, following the recommended quantity for this kind of procedure. A survey weight for each new household is obtained by dividing the original survey weight by three.

D. Computation

This section describes the computational methods used to solve the quantitative model in Section 4. We need to (i) solve a household problem with a continuum of housing assets with different service flows and prices and (ii) solve for the equilibrium objects (service flow for 2000, and price for 2005) given the three-dimensional distribution of household characteristics and the one-dimensional distribution of house qualities.

Both the price and service flow functions are defined on the interval of available house qualities $[\underline{h}, \overline{h}]$. Both functions are parametrized as shape-preserving cubic splines, defined by a set $\{h_i, s_i, p_i\}_{i=1}^{I}$, where $h_i \in [\underline{h}, \overline{h}]$ are the break points, $p_i \in [0, \infty)$ is the price p_i at h_i and s_i is the service flow at h_i. We impose strict monotonicity on both functions, that is, $h_j > h_i$ implies $p_j > p_i$ and $s_j > s_i$. Denote the approximating price and service flow functions by $\widehat{p}(h)$ and $\widehat{s}(h)$, respectively. The intertemporal household problem is tractable even with a continuum of assets because agents expect permanent shocks to not alter relative prices across houses. The price function expected in the future equals the cumulative permanent innovation to house prices plus the price function $\widehat{p}(h)$.

To accurately capture the covariation in the three mover characteristics (age, income and wealth), we use the distribution derived from the Census and SCF using the imputation procedure in Appendix C. For every survey household i at date t, we have a tuple (a_{it}, y_{it}, w_{it}) as well as a survey weight. We solve the household problem for every survey household i and obtain

his preferred house quality. We then use the survey weights to construct a cumulative distribution function for house quality. In equilibrium, this cdf must be equal to the house quality cdf from the data, shown in Figure 2. The equilibrium object (price or service flow) is found by minimizing a distance between those cdfs.

Household problem

The solution to the household problem is calculated using finite-horizon dynamic programming. Value and policy functions are approximated by discretizing the state space on a fine grid. Consider the optimization problem faced by a household of age a, with cash w as defined in equation (16), income y, and house of quality h. Each period, the household receives an exogenous mobility shock: $m = 1$ indicates that the household must move and $m = 0$ otherwise. The vector of state variables at time t is

$$x_t = [a_t, y_t, w_t, m_t, h_t].$$

The value function at time t is denoted $v(x_t)$. Income is a separate state variable even though the only shocks to income are permanent. This is because house prices are hit by shocks other than income shocks – the common approach of working with the wealth/income ratio and house/income ratio as state variables does not apply.

It is helpful to separate the household's moving decision from the other choices he makes conditional on moving or staying. Consider first a household who is moving within the period. He decides how to allocate cash on hand (which could come from a prior sale of a house) to consumption, housing or bonds, subject to the budget and collateral constraints. Denote the "mover value function" for this problem by v^m; it depends on the state as well as the approximating price and service flow functions. Consider next a household who is staying in a house of quality h. He decides how to allocate cash on hand to consumption or bonds, again subject to budget and collateral constraints. A stayer household is thus allowed to change his mortgage – this assumption is appropriate for the boom period where refinancing and home equity loans were common. Denote the stayer value function by v^s.

A homeowner who lives in a house of quality h and who does not have to move ($m = 0$) has the option of either selling the house and incurring

the transaction cost for selling, or staying in the same house. The selling household faces the same optimization problem as a moving household, with cash adjusted for the sales transaction costs. The optimization problem of the staying household is characterized by the stayer value function we defined above. Thus the value function of the owner household v^0 is the maximum of both options

$$v^o\left(a, y, w, h; \hat{p}, \hat{s}\right) = \max\left\{v^s\left(a, y, w, h; \hat{p}, \hat{s}\right), v^m\left(a, y, w, h; \hat{p}, \hat{s}\right)\right\}.$$

The beginning-of-period value function v takes into account both forced moves ($m = 1$) and endogenous moves:

$$v(x) = m\, v^m(a, y, w; \hat{p}, \hat{s}) + (1 - m)\, v^o\left(a, y, w, h; \hat{p}, \hat{s}\right)$$

We specify separate approximating functions for $v^m(\cdot)$, $v^o(\cdot)$, as well as the housing policy function $h^m(a, y, w; \hat{p}, \hat{s})$ associated with $v^m(\cdot)$. Since the downpayment constraint may induce kinks in the cash dimension w, we perform a discrete approximation of both functions separately for each age. In particular, for each age a, we specify a two-dimensional grid in (y, w)-space for the mover function $v^m(\cdot)$, and a three-dimensional grid in (h, y, w)-space for the stayer function $v^s(\cdot)$. We then maximize the value function at each grid point by searching over the set of all feasible choices at that point. To capture the effect of the downpayment constraint as precisely as possible, the grid in the cash dimension has a higher density of points for low levels of cash.

Specifically, we use 25 grid points each for cash and income dimensions, and 175 grid points for the housing dimension of the owner value function. For income and cash, 15 equally spaced points are used to cover the interval between $0 and $600K, and the remaining 10 points are equally spaced between $600K and the upper bounds of the grids. The upper bounds are set to the 98th percentile of the respective distribution in the data. Policy functions are linearly extrapolated for those observations above the upper bounds.

The discrete choice of moving in combination with the transaction cost of selling introduces a "region of inaction" for the owner households. In absence of the transaction cost, the mover and owner value functions would be

9

identical, since we have defined cash to include the sales price of the house,. Thus, without selling frictions, there is a unique optimal house quality choice $h^m(a, y, w; \hat{p}, \hat{s})$ for each level of cash and income. In the presence of transaction costs, however, the homeowner may decide to stay in a house of quality different from the frictionless optimum $h^m(a, y, w; \hat{p}, \hat{s})$. This is the case if the endowed house h is not too far from $h^m(a, y, w; \hat{p}, \hat{s})$ given the size of the transaction cost. More precisely, for a given level of transaction cost ν, there exists an interval $\left[\underline{h}(a, y, w; \hat{p}, \hat{s}, \nu), \overline{h}(a, y, w; \hat{p}, \hat{s}, \nu)\right]$ around the optimal mover choice $h^m(a, y, w; \hat{p}, \hat{s})$, such that – if the endowed quality h is located within this interval – the owner will optimally stay in the current house. Intuitively, if the endowed quality h is close to the frictionless optimum, then saving the transaction cost outweighs the benefit of adjusting the consumption bundle over housing and other goods.

We use this structure of the problem to efficiently compute the owner function $v^o(a, y, w, h; \hat{p}, \hat{s})$. We first solve the mover problem by searching over all feasible choices for each combination of age a, income y and cash w to obtain $h^m(a, y, w; \hat{p}, \hat{s})$. We then find the bounds of the inaction interval $\left[\underline{h}(a, y, w; \hat{p}, \hat{s}, \nu), \overline{h}(a, y, w; \hat{p}, \hat{s}, \nu)\right]$ for each combination of age a, income y, and cash w. We do this by checking for different house quality levels whether a household with characteristics (a, y, w) who is endowed with quality h would prefer to stay in house h instead of selling the current house and buying the mover optimum $h^m(a, y, w; \hat{p}, \hat{s})$. We start this search at the mover optimum $h^m(a, y, w; \hat{p}, \hat{s})$, and then search on a fine grid upwards and downwards from this point in the house quality space. In either direction, once we have a found a house quality level at which selling and moving is preferred over staying, we know that we have found the bound of the inaction region. This procedure minimizes the number of optimization problems we need to solve to compute the owner value function.

Figure 12 plots the policy function for a young mover (aged 28 years). Since age a is fixed, the policy function depends on the income-cash ratio y/w and cash w. Higher y/w ratios mean a larger share of human wealth out of total wealth. Movers with higher y/w ratios choose a riskier portfolio with more housing. Housing demand is also increasing in cash w. In a dynamic programming problem with collateral constraints and linear pricing, we expect the policy function to be convex in the constrained region and then to become linear for higher income-cash ratios and cash levels. Figure
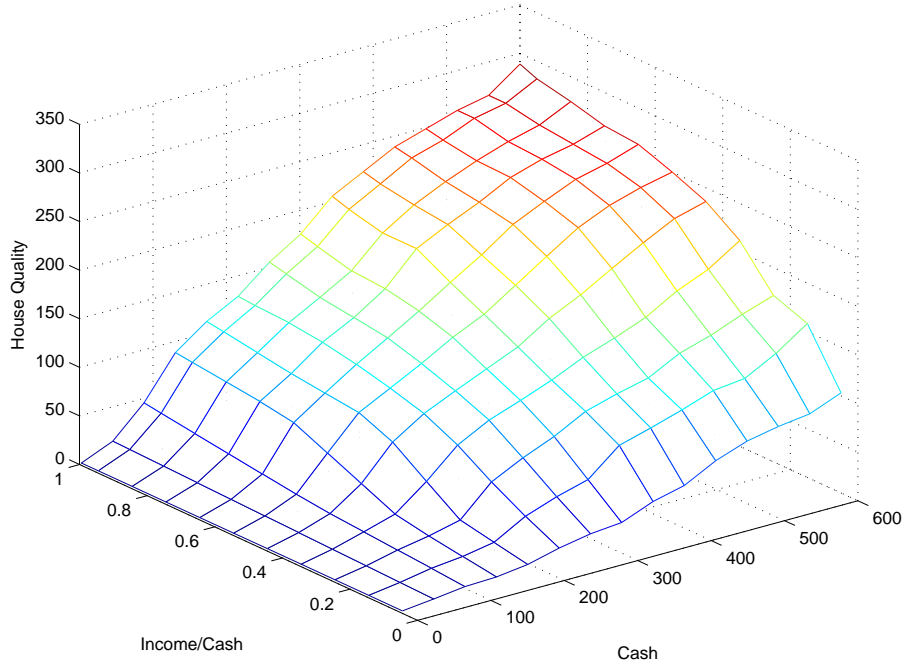
Figure 12: Policy function for young mover

12 shows more curvature throughout the state space, because the price of house quality is nonlinear.

Older movers have policy functions that look qualitatively similar to Figure 12. The older movers choose lower house qualities, especially at low income-cash levels and overall cash levels. Moreover, their policy function reaches the unconstrained region sooner, i.e. at lower values of the two state variables.

Market clearing

Given a sample of movers with characteristics $\{a_{it}, y_{it}, w_{it}\}_i$ as well as approximating price and service flow function $\hat{p}$ and $\hat{s}$, we calculate the model-implied optimal house qualities as

$$\widehat{h}_{it} = h^m \left(a_{it}, y_{it}, w_{it}; \widehat{p}, \widehat{s} \right).$$

We thus obtain a sample of optimal house quality choices $\{\widehat{h}_{it}\}_i$. We then use the survey weights for the movers to compute an empirical cdf of house

11

quality. We smooth this cdf using a cubic spline. We call the resulting cdf $\hat{G}^{dem}(h; \widehat{p}, \widehat{s})$ the demand cdf as it represents optimal housing demands at the given price and service flow functions.

In equilibrium, the demand cdf must equal the quality cdf from the data. The latter is also given as a cubic spline, $\hat{G}$ say, as explained in Appendix C. To get a measure of distance between the demand cdf and the data quality cdf, we define a set of test quantiles $\{g_j\}_{j=1}^{N_G}$, $g_j \in (0, 1)$ and compute

$$\sum_{j=1}^{N_G} \left\{ \hat{G}(\left[\hat{G}^{dem}\right]^{-1}(g_j; \widehat{p}, \widehat{s})) - g_j \right\}^2. \qquad (A\text{-}1)$$

For our exercises we need to find the equilibrium object (price or service flow), taking as given the respective other function (service or price). In each case, our algorithm chooses the spline coefficients of the equilibrium function to minimize the distance (A-1). For the reported results we use 7 break points and the test quantiles are the nine deciles between 10% and 90% as well as the 5-th and 95-th percentiles.

Figure 13 shows that the errors are within one percentage point at every test quantile. The errors labeled "service" are for the results in Figure 15. The "2005 model" errors are for the results in Figure 8. The "only house density" and "only credit conditions" errors are for the two models in Figure 10, respectively. The "const. exp. cap gains" errors are for Figure 14.

E. Sensitivity to assumptions on expectations and service flow

This appendix checks the sensitivity of our results with respect to our assumptions on expectations and the service flow function. Our model assumes that households expect favorable credit market conditions (in particular, low downpayments and spreads) to remain in place and house prices to grow at trend. While such expectations are consistent with survey evidence for the peak of the boom in 2005, they were of course disappointed during the Great Recession. To examine the sensitivity of our results to expectations of future market conditions, we thus consider a "perfect foresight" scenario designed to capture recent developments in housing and financial markets.
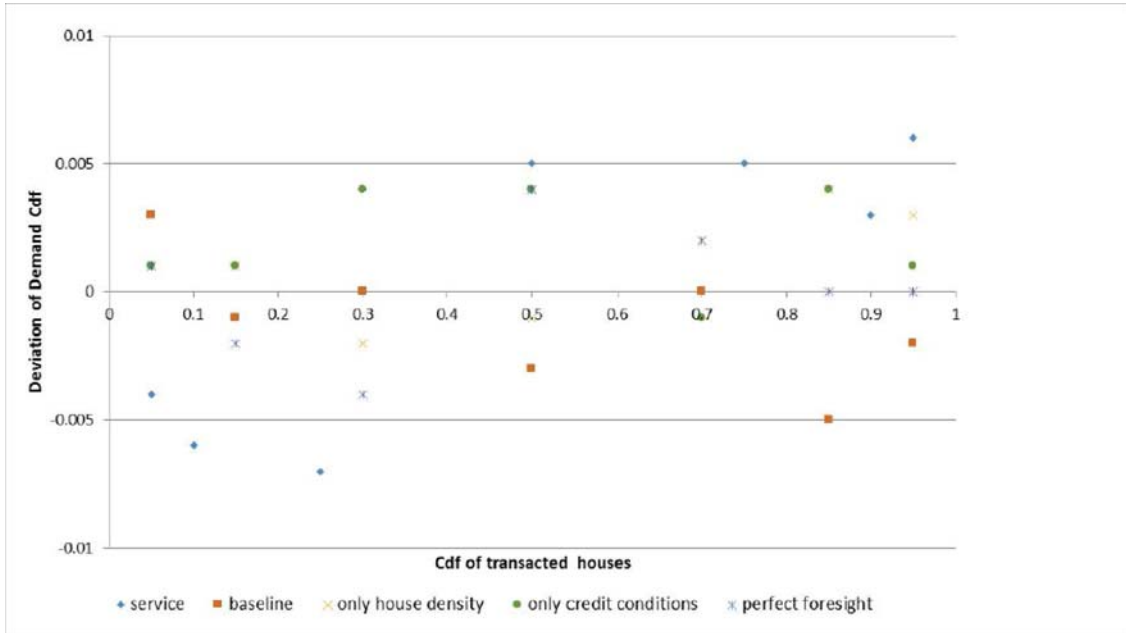
Figure 13: Errors in market clearing conditions

The perfect foresight scenario retains 2005 mover and house distributions. However, households now expect that, after three years (or one model period), (i) downpayment constraints and mortgage spreads to return to their 2000 benchmarks after three years, (ii) house prices return to 2000 prices plus trend after three years, (iii) the volatility of idiosyncratic shocks to housing returns increases to 11.8% from 9%. Households also expect that (iv) interest rates remain permanently low at 1%. Figure 14 displays equilibrium capital gains under perfect foresight.

The results show that if households had perfectly foreseen conditions of the housing bust in the Great Recession, the house price boom would have been substantially smaller. However, the boom would have generated the same cross sectional patterns in capital gains: for low quality houses, 2000-5 capital gains rise above 10% per year, about 40% of the total observed gain. This happens even though capital gain expectations under perfect foresight are actually worse for low quality houses. For low quality buyers, current favorable credit conditions thus outweigh pessimistic expectations of future market developments.
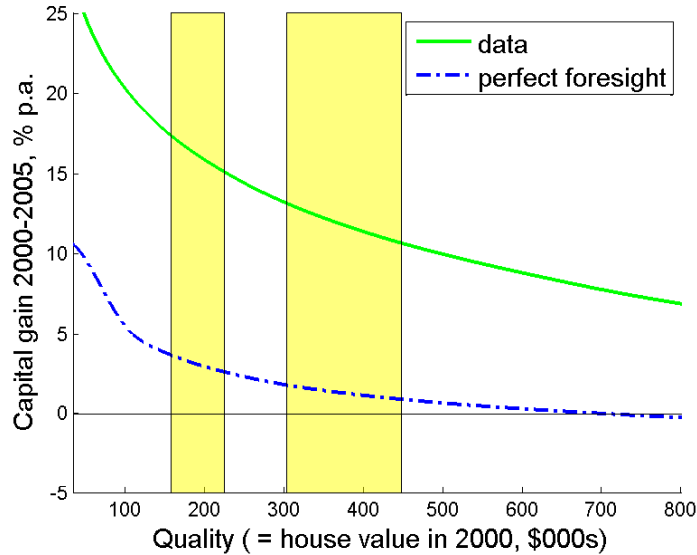
Figure 14: Model Results for 2005 under perfect foresight scenario. The green line represents capital gains in the data, while the dotted blue line shows the model counterpart.

Our model also assumes that the service flow function remains the same over time, up to a growth factor. To check the importance of credit conditions under alternative assumptions on this function, we compute a new service flow function that exactly matches 2005 observed house prices. This computation also uses (i) 2005 distributions for house qualities and mover characteristics, (ii) 2005 credit conditions, and (iii) constant capital gain expectations. The left panel of Figure 15 compares the benchmark (blue) and the service flow function that exactly matches 2005 house prices (green.) The new service flow function grows faster for qualities below 400K, which helps match 2005 prices. The right panel of Figure 15 shows the resulting capital gains. By definition, the model-implied capital gains are identical to those in the data.

To again isolate the importance of credit conditions, we now recompute the model with the new service flow function but under 2000 credit conditions. The result is the dashed line in the right panel of Figure 15. It can be compared to the right hand panel in Figure 10 which also displays a change in

14

credit conditions alone. In both cases cheaper credit increases capital gains by similar magnitudes.
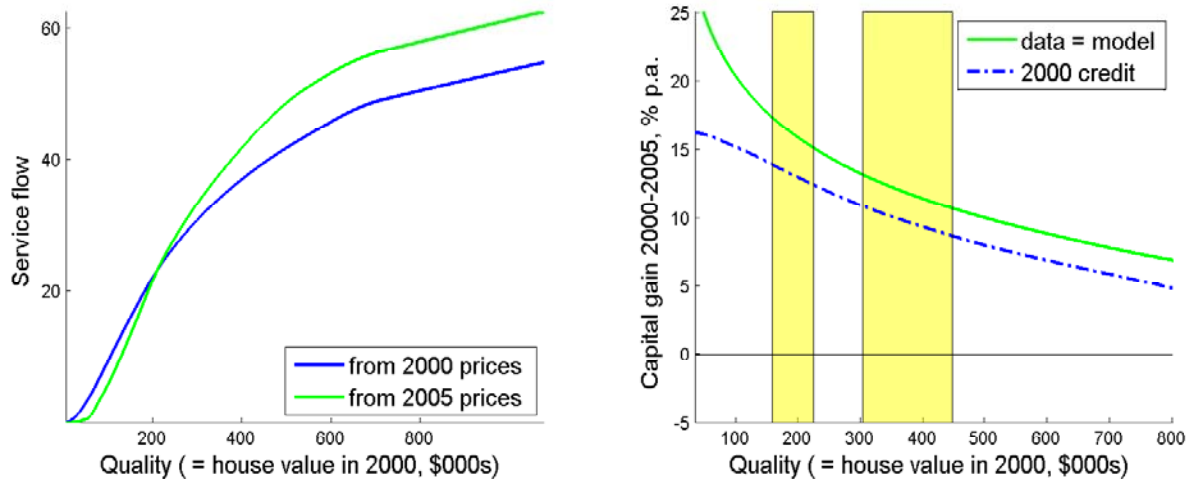

Figure 15: The left panel plots two service flow functions. The blue line matches 2000 house prices as in the benchmark. The green line is a new service flow function that matches 2005 house prices. The right panel shows capital gains under the new service flow function, which are identical to the data. The dashed line computes equilibrium prices with the new service flow function and 2000 credit conditions.