On the Nature of Self-Assessed House Prices*

Morris A. Davis† Rutgers University

Erwan Quintin‡ University of Wisconsin

June, 2016

Abstract

In models of optimal household behavior, the value of housing affects consumption, savings and other variables. But homeowners do not know the value of their house for certain until they sell, so while they live in their home they must rely on local house price data to estimate its value. This paper uses data from the recent housing boom and bust to demonstrate that changes in households' self-assessed home values are strongly consistent with the predictions of a model in which households optimally filter available house price data. Specifically, we show that self-assessed house prices did not increase as rapidly as house price indexes during the boom and did not decline as severely during the bust. A Kalman Filter model nearly perfectly replicates these data. These findings have direct implications for economists studying asking prices during booms and busts, optimal default decisions and other key housing-related phenomena.

*We thank Colin Caines, Gabe Ehrlich, Jim Follain, Carlos Garriga, Andra Ghent, Richard Green, Albert Marcet, Steve Malpezzi, Norm Miller, Andy Paciorek, Carles Vergara-Alert, Tony Yezer and seminar participants at a number of conferences and institutions for comments and suggestions.

†Department of Finance and Economics, Rutgers Business School, Rutgers University, 1 Washington Park #1092, Newark, NJ 07102; morris.a.davis@rutgers.edu; `http://morris.marginalq.com/`.

‡Department of Real Estate and Urban Land Economics, University of Wisconsin-Madison, 5257 Grainger Hall, Madison, WI 53706; equintin@bus.wisc.edu; `http://erwan.marginalq.com/`.

1 Introduction

While housing accounts for much of the wealth in the United States, the exact market value of any given house is only known at the time of a sale. As a result, homeowners are unsure of the value of the largest and most important asset in their portfolio. There are several well understood causes of this uncertainty. First, while homeowners and appraisers can observe sales prices of nearby similar homes ("comps"), any differences in location and other attributes imply that comps are imperfect benchmarks, at best.[1] In addition, housing markets are characterized by search and matching frictions so that even identical homes may sell for different prices.[2]

Given households must guess at the current value of their home, a number of papers have explored whether self-assessed home values are biased.[3] These studies are of primary importance to many Economists including those that study household saving and portfolio decisions over the life cycle. In nearly all models in this broad literature, it is typically assumed that households maintain an unbiased estimate of the current value of their home.[4] The research to date suggests that this assumption is reasonable: Homeowners, on average, tend to report accurate estimates.

Our paper also studies how households evaluate the current market value of their house but focuses on how households update their estimates of market value when surrounding house prices and area house price indexes are changing rapidly. To be specific, we study the

[1] According to their detailed analysis of pricing characteristics of 59 metropolitan areas, Malpezzi, Ozanne, and Thibodeau (1980) find hedonic regressions of house prices on housing attributes typically yield R^2 values in the range of 0.50 to 0.75. Even with more localized data on sales prices, variation in house prices after controlling for observed characteristics is still a prominent feature of the data. The web site Zillow, for example, lists confidence intervals for its ability to predict the sales price of any home given the sales price of nearby homes. These intervals can be large: See the table under the heading "Data Coverage and Zestimate Accuracy," at the web site *http://www.zillow.com/zestimate/#what.*

[2] In many models of search and matching, an ex-post distribution of sales prices arises for an ex-ante identical set of homes. A classic example is Yavas (1992) who describes a an environment where sellers and buyers with private valuations meet via random search and trading prices are determined via Nash-bargaining. See Han and Strange (2015) for a recent review of the housing-search literature.

[3] For example, see Follain and Malpezzi (1981), Kiel and Zabel (1999), Bucks and Pence (2008), Kuzmenko and Timmins (2011) and Genesove and Han (2013).

[4] More precisely, most models in finance and macroeconomics assume households know their current (unbiased) house value with certainty: See Davis and Van Nieuwerburgh (2015). Recent notable exceptions are by Ehrlich (2014) and Corradin, Fillat, and Vergara-Alert (2015).

behavior of self-assessed home prices during the recent housing boom and bust. Our key finding is that homeowners do not immediately adjust self-assessed house prices by the full change in local house price indexes. Rather, they gradually update their assessments, such that self-assessments nearly fully reflect changes in house price indexes only after some time has passed. This is true for both the boom and the bust.

We show that a simple optimal learning framework can account for almost all of the disconnect between the change in self-assessed house prices and the change in local house price indexes. Given that house prices are subject to random shocks and that homeowners only observe the set of comparable house prices (i.e. a "noisy" signal of the current value of their homes), it can be shown that homeowners should optimally apply a Kalman Filter to update their guess of the market value of their home. The formula for this optimal update is a weighted average of two pieces: (i) their previous self-assessed house price and (ii) the noisy signal on current house prices, i.e. the sale prices of nearby relevant homes. The Kalman Filter model has the desirable property that in the long-run self-assessed house prices are unbiased on average, consistent with prior findings, but it also easily explains the sluggish and lagging response of self-assessments to large changes in local price indexes.

We use a maximum likelihood procedure to estimate the parameters of our Kalman-Filter model after merging publicly available data on average self-assessed house prices from the Census and American Community Surveys and house price indexes from Case-Shiller-Weiss (CSW) for 20 metropolitan areas. Using only year-2000 self-assessment data, we use our model to generate rolling out-of-sample predictions of the annual average of self-assessed house prices in these 20 metro areas from 2006-2011. A regression of actual on predicted self-assessed house prices produces an $R2$ value of 0.97. These findings strongly suggest that during periods of rapidly changing house prices households optimally filter the available data.

Our research has fundamental implications for the way one might interpret several established findings in the housing literature. First, it is well documented that appraisals tend to be backwards looking and smoother than transactions prices,[5] the so-called "appraisal

[5]See Geltner (1991) and references therein.

bias." If appraisers optimally apply a Kalman Filter to noisy sales data to estimate the level of house prices, appraisals will display precisely those characteristics. Second, our Kalman Filter model could explain the observed variation in the asking price of near-identical housing units, as documented in Genesove and Mayer (2001). Since optimal filtering implies people are slow to take on changes in house price indexes, recent purchasers are more likely to set asking prices closer to their own purchase price than homeowners with longer tenure. Third, our results might explain why listing behavior and time-on-market change during a boom-bust cycle. Our model predicts self-assessed house prices will be relatively low during a boom and high during a bust. To the extent that asking prices reflect self-assessments, the duration of time on market should therefore be shorter and the frequency of multiple-bid sales should be greater during a boom as compared to a bust, consistent with the available data.

Finally and perhaps most importantly, the gap between self-assessed values and public estimates of home prices might help explain the fact that most households with negative home equity choose not to default on their mortgages (Gerardi, Shapiro, and Willen, 2009). As the default literature has discussed at length, households with negative equity often choose to continue making mortgage payments; the presumption is that households make these payments because default triggers tangible and intangible costs and because home prices could improve in the future. Our results suggest a complementary explanation. In empirical default models, researchers estimate current home value by applying the change in a house price index to the purchase price of each home. As we document, during a housing bust self-assessed house prices do not decline by as much as house price indexes. Households will self-assess greater home equity (or less negative equity) than traditional estimates.

In the next section, we report some basic facts about self-reported house prices and house price indexes during the housing boom and bust. In the third section we derive the Kalman Filter model and likelihood we use to estimate model parameters. In the fourth and fifth sections we analyze model fit and derive some other implications of our model. We then conclude.

2 Overview of the Data

In this section we introduce and compare our data main sources of data. Throughout the paper we only work with inflation-adjusted data. For 20 Metropolitan Statistical Areas (MSAs) over the 2000-2011 period, we compare real changes to the Case-Shiller-Weiss (CSW) house price index to real changes to self-reported house prices that we compile from survey data from the Decennial Census of Housing and the American Community Survey. Relative to other data sources with housing information such as the American Housing Survey and the Survey of Consumer Finances, the sample sizes for the Census and American Community Survey data are extremely large as shown in table 1. The bottom row of the table reports the percentage of the sample where self-reported house value is top-coded after accounting for sampling weights. This percentage is small enough that we do not adjust for top-coding in our data work. In the appendix, we document details of the data construction. The most important takeaways are that we do not use any imputed data on house prices from the Census and American Community Survey; and, the CSW indexes and the self-reports cover roughly the same geography and set of homes in each of the MSAs we study.

To give a sense of patterns in the raw data, Table 2 reports the real ($2005) average self-reported value of housing by year and metropolitan area. The table shows there is meaningful variation in the average of the self-reports across MSAs and over time. The median of the MSA-average values increases from $209 to $345 thousand during the housing boom and then falls to $246 thousand by 2011. In 14 of the 20 MSAs, the average of the self-reported values peaks at the same time or one year after the CSW index peaks.[6]

Over our sample period the CSW indexes and the self-reported data do not move in lock-step, the focus of this paper. Figure 1 compares the CSW index (solid line, "CSW") to the average of the self reports (dashed line, "SR") for the Los Angeles MSA; the other metro areas in our sample have similar patterns, as we show later. The SR line is linearly interpolated for the years 2001-2004, as no self report data exist in that period, and the average of the self-reports and the CSW indexes are each normalized to 100 in the year in

[6]In the case of Dallas, we adjust the real peak date of the CSW from 2002, a year lacking self-report data, to 2006. This reduces the peak value of the CSW by 1.3 percent.

which the CSW peaks, 2006 in the case of Los Angeles. Figure 1 shows that before the peak, the self-reports do not increase by as much as the CSW price index;[7] and, after the peak the CSW declines more sharply than the self-reports. In many MSAs, the difference in the percentage decline of the CSW index and the self reports during the housing bust is quite large. Table 3 reports the percentage decline in the average self-report data and the CSW indexes, measured from the MSA-specific peak date of the CSW to 2011. At the median MSA, the CSW declines by 14 percentage points more than the self-reports. In the extreme case of the Los Angeles MSA, the CSW index declines by 44 percent whereas the self-reports only decline by 18 percent.

The magnitude of the difference between the CSW and self-reported values is strongly correlated with the change in the CSW index. This is true for both the boom and bust periods. The top panel of figure 2 shows results from the boom period. The figure shows 20 dots, one for each MSA. For each MSA, cumulative real growth in the CSW index during the boom, the x-axis, is plotted against cumulative real growth in the self-reports. Figure 2 shows that growth in the CSW and self-reports are highly correlated but growth in self-reports does not reflect growth in the CSW index on a percentage point-for-percentage point basis. When cumulative growth in the self-reported data over this period is regressed on cumulative growth in the CSW index, the $R2$ of the regression is 0.95 but the coefficient is 0.74 with a standard error of 0.04 and the intercept is 15.9 with a standard error of 3.0. If the self reports tracked the CSW, we would expect an intercept of 0 and a coefficient of 1.

The bottom panel of figure 2 is constructed in an identical fashion as the top panel, except it shows results from the bust period, the CSW peak date through 2011. As with the boom period, cumulative growth of the two series are highly correlated, but they do not move percentage point-for-percentage point. A regression of cumulative growth in the self-reported data over this period on cumulative growth in the CSW index yields a coefficient of 0.86 with a standard error of 0.09; an intercept of 9.4 with a standard error of 3.5; and a $R2$ value of 0.85. In this case, we cannot reject the hypothesis that the coefficient is 1.0. However, during the bust period the regression estimates are sensitive to the ending year of the sample.

[7] Additionally, in Los Angeles the self-reports continue to rise slightly until 2008.

When the sample is specified to end in 2009, the same regression yields a coefficient of 0.58 with a standard error of 0.08 and an $R2$ of 0.74; and with an end date of 2010, the coefficient is 0.74 with a standard error of 0.08 and an $R2$ of 0.82. The increase in the coefficient from 0.58 with a sample end date of 2009 to 0.86 with an end date of 2011 – a result that shows that over time, the self-reports gradually adjust to the CSW index – supports our Kalman Filter model that we describe in the next section. The basic intuition from the model is that when information about house prices is imperfect, homeowners slowly revise down their self-assessed home value with each passing year the CSW remains depressed.

In summary, during both the housing boom and bust, the averaged self-report data do not increase or decrease percent-for-percent with changes to the CSW indexes although the two series are highly correlated. We next show that these results are consistent with an environment in which homeowners receive noisy signals (such as the CSW) about the value of their home and optimally filter these signals when determining their home's current value.

3 A Kalman Filter Model

3.1 Specification

We start by specifying a process for true but unobserved log real house prices during the housing boom and bust. Denote the true but unobserved log real price in period t of the home owned by homeowner i as h^*_{it}. We assume that this true house price follows a first-order autoregressive process with autoregressive parameter ρ and shock e_{it}.

$$h^*_{it} \quad = \quad \bar{h}_i + \rho h^*_{it-1} + e_{it} \ . \tag{1}$$

In the event $\rho < 1$, the fixed mean of this process for homeowner i is $\bar{h}_i / (1 - \rho)$. If $\rho = 1$, this process is a random walk and the true level of house prices has no fixed mean.

Homeowner i does not directly observe h^*_{it}; after all, as we noted in the introduction, the price of a home is only directly observed at the time of a sale. Instead, homeowner i observes a noisy but unbiased signal of the true log price. Denote this signal as h_{sit} such that

$$h_{sit} = h^*_{it} + \nu_{it} \ . \tag{2}$$

To make progress, we assume the shock to growth rate of prices, e_{it}, and the measurement error on the signal of the level of prices, v_{it}, are independently drawn from each other and over time. Further, we assume e_{it} and v_{it} are Normally distributed with mean 0 and variances σ_e^2 and σ_ν^2, respectively. Although e_{it} and ν_{it} are independent of each other, we allow e_{it} and e_{jt} to be correlated and ν_{it} and ν_{jt} to be correlated for any two homeowners i and j in the same metro area at the same time.

Now denote homeowner i's self-assessed value of the house as of date $t-1$, her "belief" about house value, as h_{bit-1}. Given the assumptions we have made, homeowner i should *optimally* update her belief in period t using a Kalman Filter that has the form,

$$h_{bit} = (1-\kappa_{it})\left[\bar{h}_i + \rho h_{bit-1}\right] + \kappa_{it} h_{sit} \ . \tag{3}$$

Notice homeowners should be sluggish to adjust to new signals. Households optimally update their belief about the current price of their home as κ_{it} times the current signal of market prices plus $(1-\kappa_{it})$ times last period's belief of the house price after appropriately adjusting for any expected autocorrelation in house prices.

κ_{it} is known as the "Kalman gain," which is a number between 0 and 1. In the event the signal is very informative, and κ_{it} is close to 1, then households' beliefs track the signal closely. If the signal is noisy, then κ_{it} will be significantly less than 1. The Kalman gain is updated each period using the following recursion (Hamilton, 1994)

$$\begin{aligned}
\mathcal{V}^p_{it} &= \rho^2 \mathcal{V}_{it} + \sigma_e^2 \\
\kappa_{it} &= \frac{\mathcal{V}^p_{it}}{\mathcal{V}^p_{it} + \sigma_\nu^2} \\
\mathcal{V}_{it+1} &= (1.0 - \kappa_{it})\, \mathcal{V}^p_{it}
\end{aligned} \tag{4}$$

Note that we would expect $\kappa_{it} = 0$ in the year that a house is purchased.[8] Starting off from this point, for any $0 < \rho \leq 1$, κ_{it} converges monotonically over time to a fixed value. The rate of convergence depends on the variances σ_e^2 and σ_ν^2.

3.2 Likelihood

To bring this model to the data we have on hand, we assume that κ_{it} has converged to its steady-state value κ for each homeowner in our sample. We discuss the plausibility of this assumption later.[9] Given this, we rewrite equation (3) as

$$h_{bit} \quad = \quad (1-\kappa)\left[\bar{h}_i + \rho h_{bit-1}\right] + \kappa h_{sit}. \tag{5}$$

Since (5) holds for any homeowner i, it holds for the average of all homeowners $i = 1, \ldots, N$. Denote the cross-sectional average of a variable in a given metro area at time t using a capital letter and a subscript m, for example $H_{bmt} = \left(\sum_{i=1}^{N} h_{bit}\right)$ for all the $i = 1, \ldots, N$ homeowners in metro area m. After taking averages and substituting notation, equation (5) can be written as an expression in MSA-level cross-sectional averages

$$H_{bmt} \quad = \quad (1-\kappa)\left[\bar{H}_m + \rho H_{bmt-1}\right] + \kappa H_{smt} \tag{6}$$

To reiterate, the assumption that the kalman gain has converged for everyone in our sample to a common number κ enables us to jump from equation (3), a statement about individual behavior and beliefs, to equation (6), a statement about cross-sectional averages inside a metropolitan area.

There are only two variables in equation (6), the average of the beliefs of the log sale price of each individual's home in a metro area, H_{bmt}, and the average of the signal of log house prices in that metro area, H_{smt}. Assuming people's self-reported house prices are

[8] Referring to the recursion equation, this occurs because the variance around self-assessed beliefs, V_{it}, is equal to 0 in the year of the sale as the sale price is exactly known at that moment.

[9] To preface our results, when we eliminate households from our data for whom it is least likely that κ_{it} has converged to its steady state value, our parameter estimates do not change.

the same as their beliefs about the sale price of their house, we observe H_{bmt} from the Census and American Community Survey.[10] Even if at the individual level the self-reports measure beliefs with classical measurement error, the sample sizes (reported in table 1) used to compute averages are so large that for each metro area in any given year, the unit of analysis, the measurement error for the average will be essentially zero.[11]

We do not, however, observe any individual house price signals that we can aggregate to the metro level. Instead, we observe the log of the CSW house price index. So, we assume the log CSW, denoted $\mathcal{H}_{mt}$, is an unbiased estimate of the average signal H_{smt} up to an metro-specific additive scale factor α_m and Normally distributed error u_{mt}

$$\mathcal{H}_{mt} \quad = \quad H_{smt} - \alpha_m + u_{mt} \ . \tag{7}$$

The scale factor is required to rescale the CSW price index to a level. Now insert (7) into (6) to get

$$H_{bmt} \quad = \quad a_m \ + \ (1-\kappa)\,\rho H_{bmt-1} \ + \ \kappa\mathcal{H}_{mt} \ - \ \kappa u_{mt} \tag{8}$$

$$\text{where} \quad a_m \quad = \quad (1-\kappa)\,\bar{H}_m \ - \ \kappa\alpha_m \ . \tag{9}$$

Notice that it looks like we can run a simple fixed-effects regression using equation (8) to uncover estimates of κ and ρ, assumed the same for every metro area, and α_m (one for each metro area).[12] However, that is not the case as the error term in equation (8), u_{mt}, is correlated with the regressor $\mathcal{H}_{mt}$ via equation (7).

[10] H_{bmt} is the average of the log of self-reported house prices. To be clear, this is not equal to the log of the average self-reported house prices that are shown in table 2.

[11] To give an example, suppose the standard deviation of reporting error at the household level in logs is 0.05, i.e. 5 percent. With a sample size of 5,000, the standard deviation of the average of the reporting error will be 0.071 percent.

[12] In the case of κ, this is equivalent to saying that σ_e^2 and σ_ν^2 are identical for all metro areas during the housing boom and bust. Obviously some metro areas had larger booms and busts than others, but it is not clear if it is preferable to specify different variances of shocks, i.e. different values of σ_e^2, or simply different realizations of shocks from the same underlying distribution. We chose the latter assumption: Many of the areas with the greatest boom and bust, like Las Vegas, had what appeared to be plenty of easily developable land and thus should not be subject to big swings in house prices (Davidoff, 2013).

To make progress, we simply rearrange (8) as

$$\mathcal{H}_{mt} \;=\; -a_m \;+\; \left(\frac{1}{\kappa}\right) H_{bmt} \;-\; \left(\frac{1-\kappa}{\kappa}\right) \rho H_{bmt-1} \;+\; u_{mt} \tag{10}$$

We estimate (10) over the 2006-2011 period (120 observations) using non-linear least squares with fixed effects. The $R2$ of the regression is 0.96 and the standard deviation of the error term is 0.0482. Coefficient estimates and standard errors are reported in the top two rows of table 4. We estimate $\rho = 1.04$ with a standard error of 0.038, implying we cannot reject house prices are a random walk, and we rather precisely estimate the kalman gain to be 0.53 with a standard error of 0.016.

One criticism observes that no information from the housing boom is used to inform estimates because the first year in the estimation sample is 2006. We therefore re-estimate model parameters using a simulated maximum likelihood procedure that takes advantage of all available data. Rewrite equation (8) as

$$u_{mt} \;=\; \mathcal{H}_{mt} \;-\; \kappa^{-1} \left[H_{bmt} - a_m - (1-\kappa)\, \rho H_{bmt-1}\right] \tag{11}$$

Denote θ as the full vector of parameters and $\ell\,(\theta)_{mt}$ as the log likelihood of the data for metro area m at year t. This log likelihood is the log of the density of u_{mt} from equation (11). Our estimate of θ maximizes

$$\mathcal{L}\,(\theta) \;=\; \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{m=1}^{20} \tilde{\ell}\,(\theta)_{mt=2005}\right] \;+\; \sum_{m=1}^{20} \sum_{t=2006}^{2011} \ell\,(\theta)_{mt} \tag{12}$$

where N is the total number of simulation draws and $\tilde{\ell}$ denotes a simulated log likelihood. The first term on the right-hand side of equation (12) shows we simulate the log likelihood for the year 2005. The values of u_{mt} for $2006-2011$ are directly observable with data on hand from equation (11) and do not depend at all on the simulations, explaining the rightmost term of equation (12).

We use simluations because there is a gap in obseved values of H_{bmt}: We first observe H_{bmt} in 2000 and then annually from $2005 - 2011$. Our simulation procedure fills in this gap. We draw u_{mt} from its distribution for each of $t = 2001 - 2004$. Given this draw, and given a value of H_{bmt} for the year 2000 and values for the CSW ($\mathcal{H}_{mt}$) for $2001 - 2004$, we sequentially apply equation (8) to generate simulated values of H_{bmt}. With a simulated value of H_{bmt} in hand for the year 2004, we use equation (11) to determine u_{mt} for 2005 and compute the log likelilhood at this draw, $\tilde{\ell}_{mt=2005}$. We repeat this process $N = 25,000$ times and compute the average value of the simulated likelihoods.

We report our maximum likelihood parameter estimates and standard errors in rows 3-5 of table 4.[13] The estimates are quite similar to those we found with nonlinear least squares. We find $\rho = 0.992$ implying house prices are essentially a unit-root process; $\kappa = 0.554$, meaning people only incorporate a little more than half of a new reading of a house price signal when updating their assessment of their home value each year; and $\sigma_u = 0.0453$ implying the standard deviation of the measurement error associated with the level of the CSW index is about 4.5%. For the rest of our analysis, we use this set of parameter estimates.

4 Model Fit

To get a sense for model fit, in figure 3 we plot the CSW data $\mathcal{H}_{mt}$ against the model-implied signal H_{smt} for the years in which the signal is computable without simulation, $2006 - 2011$. In other words, given our parameter estimates, we back out values of the signal H_{smt} such that equation (6) *exactly* holds and compare this signal to the observed CSW. To ensure all series in the figure are appropriately scaled with a zero mean, we add an estimate of α_m to the CSW series and subtract an estimate of $\bar{H}/(1-\rho)$ from both series.[14] The graphs show that the gaps between the model-implied signal and the CSW indexes are relatively small given the large decline over time in the CSW indexes. A regression of the de-meaned CSW indexes on the de-meaned values of H_{smt} for all the MSAs and years shown in figure 3 yields an R^2 value of 0.95 with an intercept of nearly exactly 0 and a slope coefficient of

[13]The reported standard errors are computed as the square root of the diagonals of the inverse of the outer product of scores. For reference, the maximized log likelihood is 232.047.

[14]We set $\bar{H}/(1-\rho)$ as the average value of H_{bmt} over the 2005-2011 period, and given our maximum likelihood estimates of a_m we set α_m such that equation (9) holds.

exactly 1.0.

Figure 3 has the flavor of plotting one-step ahead prediction errors. Another way we informally evaluate model fit is by computing a full out-of-sample forecast. We ask how well the model could have predicted the sequence of average self-reports from 2005-2011 given (a) data on self-reports for *only* the year 2000 and (b) assuming the Case-Shiller-Weiss values for 2001-2011 were exactly equal to household signals. Given our estimates of a_m, a sequence of $\mathcal{H}_{mt}$, and a starting value for H_{bmt-1} (the year-2000 value, but no other values), we generate a model-predicted sequence of values for H_{bmt} for the years 2001-2011. To be crystal clear, data on H_{bmt-1} from only the year 2000 is used to generate predicted values. No other data on self-reports are used in this exercise. Given that ρ is essentially 1, the possibility for error late in the sample is large as there is nothing inherent in this exercise that pulls out-of-sample predictions of H_{bmt} towards the data.

Figure 4 shows a scatter diagram of the the self-reported home value data (y-axis) against the predicted self-reported values (x-axis) for all 20 metro areas over the years 2005-2011. We subtract our estimates of $\bar{H}_m/(1-\rho)$ from all data such that goodness of fit abstracts from across-MSA differences in the average level of self-reported values. The R^2 of the pictured regression line of the self-report data on predicted values is 0.97 and the intercept of the regression is nearly exactly zero. The point estimate of the slope coefficient is 0.96 with a standard error of 0.014, implying that the actual self-reports vary a bit less than the out-of-sample predictions.

5 Other Implications

As a final part of our analysis, we estimate the variance parameters σ_ν^2 and σ_e^2. Additional assumptions on the nature of the correlation of shocks across households in an MSA are required to estimate these parameters. To make progress, we assume that homeowners in each metro area experience identical values of e and ν – that is, $e_{imt} = e_{jmt}$ and $\nu_{imt} = \nu_{jmt}$ for all homeowners i and j in metro area m in period t – but allow values of e and ν to vary across metro areas. The assumption that all agents in a given MSA receive the same sized shocks is not innocuous and results derived in this section should be viewed appropriately.[15] On the other hand, a representative agent in each MSA is often assumed in models of Urban Economics.

With the assumption of a representation agent in each metro area, it can be shown that

$$\text{var}\left[H_{smt} - \rho H_{smt-1} - \bar{H}_m\right] \quad = \quad \sigma_e^2 + \left(1+\rho^2\right)\sigma_\nu^2 \ . \tag{13}$$

Additionally, it is possible to show that once the Kalman gain converges to its steady-state value it satisfies

$$\left[\sigma_\nu^2 \rho^2\right]\kappa^2 + \left[\sigma_e^2 + \sigma_\nu^2\left(1-\rho^2\right)\right]\kappa - \sigma_e^2 \quad = \quad 0 \ . \tag{14}$$

Although equation (14) is quadratic in κ, it is linear in σ_ν^2 and σ_e^2.

Given estimates of ρ, κ, and $\text{var}\left(H_{smt} - \rho H_{smt-1} - \bar{H}_m\right)$, equations (13) and (14) uniquely determine σ_ν^2 and σ_e^2, with the closed-form expression for σ_ν^2 as

$$\sigma_\nu^2 \quad = \quad \frac{\text{var}\left[H_{smt} - \rho H_{smt-1} - \bar{H}_m\right](1-\kappa)}{1+\rho^2(1-\kappa)^2} \tag{15}$$

and the expression for σ_e^2 naturally following from (13) and (14).

We estimate $\text{var}\left[H_{smt} - \rho H_{smt-1} - \bar{H}_m\right]$ using data across all metro areas and years (2007-2011)[16] to be 0.008. At our estimated values for ρ and κ, and given the computed values of H_{smt} and $\bar{H}$ in each period and metro area, we compute $\sigma_\nu = 0.0556$ and $\sigma_e = 0.0464$. The interpretation of these findings is that the standard deviation of shocks to home prices is 4.64 percent per year; and homeowners understand the standard deviation of the gap between the signal they receive on the value of their home and the true value of their home is 5.56 percent. Referencing notation in equation (4), these estimates imply a steady-state value of the square root of $\mathcal{V}_{it}$, the standard deviation of homeowners' uncertainty about the

[15]In fact, one of us has written a paper documenting that the magnitude of house price declines during the housing bust varied quite a bit within some metro areas (Davis, Oliner, Pinto, and Bokka, 2016).

[16]The first year we can estimate u_{mt} and thus H_{smt} is 2006.

the value of their home, of 4.14 percent. In other words, once we assume a representative agent inside each metro area we estimate a two standard error confidence interval around homeowners' current guess of house prices of ±8.28 percent.

Given estimates of σ_ν^2 and σ_e^2, we can compute the sequence of optimal Kalman gains for a homeowner starting the year she knows her true log house price with certainty.[17] In year 1, the Kalman gain is 0.410; year 2 it is 0.524; year 3 it is 0.548; in year 4 it is 0.552 percent; and so forth. With these estimates on hand, we now reconsider the impact of the assumption on our analysis that κ_{it} has converged to its steady state value for all homeowners in our sample. In the Census and American Community Survey data, between 8% (2011) and 14% (2005) of the sample has lived in their current house 2 years or less. When we exclude these households from our estimation sample, our maximum likelihood estimates of ρ and κ are essentially unchanged.

6 Conclusion

Given no two houses are exactly alike, it seems reasonable to assume that homeowners cannot learn the exact price at which their house will sell based on nearby sales of similar homes. Rather, homeowners should use available information to guess the current market price of their home. We show that when house prices are subject to random shocks and when the information people use to learn about house prices is unbiased but noisy, households optimally update the guess of the sale price of their home using a Kalman Filter. We estimate the parameters of a Kalman Filter model for house prices using data for 20 metropolitan areas during the housing boom and bust and show that house prices are essentially a random walk and the steady state Kalman gain is slightly above 0.5. Our analysis of these data strongly suggest to us that the Kalman Filter model is appropriate, in the sense that the model accurately replicates the data as shown in figures 3 and 4. Our expectation is that future researchers will use our results to better understand the nature of appraisal bias, asking prices during boom-bust cycles and optimal default decisions.

[17] For convenience, think of this as the year that she bought her home.

References

Bucks, B., and K. Pence (2008): "Do Borrowers Know their Mortgage Terms?," *Journal of Urban Economics*, 64(2), 218–233. 1

Corradin, S., J. L. Fillat, and C. Vergara-Alert (2015): "Portfolio Choice with House Value Misperception," Working Paper, IESE. 1

Davidoff, T. (2013): "Supply Elasticity and the Housing Cycle of the 2000s," *Real Estate Economics*, 41(4), 793–813. 9

Davis, M. A., S. D. Oliner, E. J. Pinto, and S. Bokka (2016): "Residential Land Values in the Washington, DC Metro Area: New Insights from Big Data," Working Paper, Rutgers University. 13

Davis, M. A., and S. Van Nieuwerburgh (2015): "Housing, Finance and the Macroeconomy," in *Handbook of Urban and Regional Economics*, ed. by G. Duranton, J. V. Henderson, and W. C. Strange, vol. 5. North-Holland. 1

Ehrlich, G. (2014): "Price and Time to Sale Dynamics in the Housing Market: the Role of Incomplete Information," Working Paper, Congressional Budget Office. 1

Follain, J. R. J., and S. Malpezzi (1981): "Are Occupants Accurate Appraisers," *Review of Public Data Use*, 9, 47–55. 1

Geltner, D. (1991): "Smoothing in Appraisal-Based Returns," *Journal of Real Estate Finance and Economics*, 4(3), 327–345. 2

Genesove, D., and L. Han (2013): "A Spatial Look at Housing Boom and Bust Cycles," in *Housing Markets and the Financial Crisis*, ed. by E. Glaeser, and T. Sinai. University of Chicago Press. 1

Genesove, D., and C. Mayer (2001): "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *The Quarterly Journal of Economics*, pp. 1233–1260. 3

Gerardi, K., A. H. Shapiro, and P. S. Willen (2009): "Decomposing the foreclosure crisis: House price depreciation versus bad underwriting," Federal Reserve Bank of Atlanta Working Paper 2009-25. 3

Hamilton, J. D. (1994): *Time Series Analysis*. Princeton University Press. 7

Han, L., and W. C. Strange (2015): "The Microstructure of Housing Markets: Search, Bargaining, and Brokerage," in *Handbook of Regional and Urban Economics*, ed. by G. Duranton, V. Henderson, and W. C. Strange, vol. 5, chap. 13. North-Holland. 1

Kiel, K. A., and J. E. Zabel (1999): "The Accuracy of Owner-Provided House Values: The 1978-1991 American Housing Survey," *Real Estate Economics*, 27(2), 263–298. 1

Kuzmenko, T., and C. Timmins (2011): "Persistence in Housing Wealth Perceptions: Evidence from the Census Data," Working Paper, Duke University. 1

Malpezzi, S., L. Ozanne, and T. Thibodeau (1980): "Characteristic Prices of Housing in Fifty-Nine Metropolitan Areas," The Urban Institute Working Paper 1367-1. 1

Yavas, A. (1992): "A Simple Search and Bargaining Model of Real Estate Markets," *Real Estate Economics*, 20(4), 533–548. 1

Data Appendix

The CSW house price indexes are derived from repeated sales of single-family housing units, which in principle delivers a constant-quality price index. We average the monthly nominal CSW index values in each year and convert the nominal annual index to real by deflating using the personal consumption price index from National Income and Product Accounts (NIPA), line 2 of NIPA table 1.1.4. The CSW discards transactions that occur with 6 months; regression weights used to compute the index are reduced the longer the time between sales; and, price anomalies are down-weighted. For further discussion, see http://www.macromarkets.com/csi_housing/documents/tech_discussion.pdf.

The data on self-assessed home values are from the 5% sample of the 2000 Census and the annual 2005 through 2011 American Community Surveys (ACS).[18] 2005 is the first year that the ACS includes metropolitan-area data. The geographic boundaries defining metropolitan areas in the 2000 Census and the 2005-2011 ACS are consistent with the 2000 definition and

[18] These data are available from the IPUMS web site, see http://usa.ipums.org/usa/. See Ruggles et. al. (2010).

are consistent with the MSA boundaries in the CSW data.[19] From the Census and ACS data, we include only nonfarm, single-family detached or attached, owner-occupied housing units. The large majority of these units are detached. We keep any housing unit where the reported house value is specified in a flag variable as "unaltered."[20] The Census and the ACS continually sample throughout the survey year. As mentioned earlier, the total number of housing units in the Census and ACS sample that meet the criteria listed above are reported in Table 1.

The Census and ACS data include information on whether or not the units are detached or attached; the age of the units; and, the number of bedrooms in each unit. We assign every unit in the sample to a bin based on these observable characteristics.[21] Attached housing units account for a small percentage of the overall sample and we bin them together by MSA. For the detached units, the bins are based on bedrooms (1 or 2, 3, or 4 or more) and age of structure: Built before 1940; decade-by-decade from 1940-1949 through 1990-1999; 2000-2004; and 2005-2011. We compute sampling weights for each bin prior to discarding any missing or imputed observations on house prices. The sum of the sampling weights across bins is 1.0 in each metro area in every year.

We calculate the average value of housing as the sampling weight for each bin multiplied by the average of the non-missing house values in that bin. We then adjust for inflation using the NIPA price index described earlier. We compute the average value of the log of real house prices using an analogous procedure.

[19]With the exception of New York and Chicago, the CSW indexes cover the full set of counties in each of the 20 metro areas. The Chicago index does not cover properties sold in the Kenosha, WI and the Gary, IN metropolitan divisions. The CSW index for New York samples all 23 counties included as part of the New York Metropolitan Statistical Area and 6 others: Fairfield and New Haven counties in Connecticut, Mercer and Warren counties in New Jersey, and Dutchess and Orange counties in New York. In 200, the excluded divisions in Chicago account for 16 percent of the MSA population and the additional 6 counties in the New York metro area increase the population by 15 percent.

[20]Of all metro areas in all years, there are 150 total instances where $0 is reported as the house value. These are treated as missing. In addition, we discard from the sample in the 2005 ACS any house reported as built in 2005. Finally, the top code for self-reported house value is $1,000,000 in 2000 and 2005-2007 and varies by state after 2008.

[21]Bins are chosen such that each bin in every metro area in every year contains at least one house price observation.

Table 1: Sample Sizes for Self-Reported House Values, 1-Family Owner-Occupied Units

2000 5% Census, 2005-2011 American Community Survey

	5% Census	American Community Survey						
MSA	2000	2005	2006	2007	2008	2009	2010	2011
Atlanta	34,865	10,986	11,180	11,462	11,130	11,398	10,579	8,637
Boston	30,588	7,952	7,875	7,873	7,546	7,706	7,596	7,572
Charlotte	13,125	4,147	4,268	4,501	4,479	4,486	4,357	3,870
Chicago	66,584	16,969	16,875	17,108	16,406	16,665	15,843	14,636
Cleveland	24,888	6,021	6,051	6,028	5,774	5,825	5,778	5,910
Dallas	42,525	12,850	12,908	13,306	13,258	13,315	12,975	12,158
Denver	20,847	5,958	5,957	6,060	5,851	5,856	5,813	5,688
Detroit	39,689	9,937	9,817	9,742	9,093	9,176	8,707	8,742
Las Vegas	11,319	3,767	3,880	3,951	3,853	3,832	3,680	3,417
Los Angeles	82,064	20,354	20,040	20,218	18,989	19,247	19,095	20,048
Miami	14,148	3,475	3,590	3,637	3,409	3,379	3,409	3,410
Minneapolis	24,213	5,883	5,943	5,895	5,806	5,730	5,489	5,161
New York	100,842	24,307	23,735	23,826	23,184	23,231	23,013	23,371
Phoenix	29,324	8,987	8,958	8,958	8,551	8,503	8,093	7,761
Portland	16,173	4,407	4,460	4,546	4,427	4,412	4,379	4,360
San Diego	20,526	5,603	5,529	5,387	5,159	5,194	5,065	5,143
San Francisco	38,864	9,583	9,411	9,371	8,780	9,052	8,870	8,774
Seattle	20,524	5,498	5,462	5,661	5,510	5,569	5,599	5,567
Tampa	24,491	7,219	7,189	7,283	6,860	6,828	6,735	7,066
Washington, DC	46,335	12,175	12,131	12,369	11,894	12,052	11,985	11,505
top-code percent	1.2	4.3	5.3	5.6	0.6	0.6	1.0	1.0

Table 2: Average of Real Self Reported (SR) House Values

2000 5% Census, 2005-2011 American Community Survey

	Thousands of $2005 Dollars								Peak Date		
MSA	2000	2005	2006	2007	2008	2009	2010	2011	SR	CSW	SR-CSW
Atlanta	193	238	245	249	248	230	207	189	2007	2006	1
Boston	310	480	475	464	455	432	423	401	2005	2005	0
Charlotte	175	205	211	217	232	219	211	204	2008	2007	1
Chicago	224	301	314	315	312	289	270	250	2007	2006	1
Cleveland	162	183	181	180	174	160	158	147	2005	2005	0
Dallas	152	178	183	187	190	188	184	178	2008	2002	6
Denver	242	308	306	304	309	299	283	277	2008	2006	2
Detroit	183	217	213	206	190	159	142	132	2005	2005	0
Las Vegas	184	355	374	369	317	236	198	176	2006	2006	0
Los Angeles	328	585	626	618	637	560	539	513	2008	2006	2
Miami	184	338	378	381	377	306	272	241	2007	2006	1
Minneapolis	189	299	297	297	285	263	242	226	2005	2006	-1
New York	296	486	508	502	510	483	457	437	2008	2006	2
Phoenix	184	292	349	341	321	261	233	200	2006	2006	0
Portland	232	293	330	348	341	316	288	274	2007	2007	0
San Diego	314	610	615	590	572	500	472	454	2006	2005	1
San Francisco	434	663	682	663	716	618	593	556	2008	2006	2
Seattle	305	398	438	466	474	438	410	373	2008	2007	1
Tampa	137	233	275	265	251	210	196	173	2006	2006	0
Washington, DC	258	471	507	500	480	439	418	399	2006	2006	0
Median	209	305	340	345	319	294	271	246	2007	2006	1
Standard Dev	76	147	151	146	154	137	134	130	1	1	1

Table 3: Real Percent change in CSW and Self Reports (SR), CSW peak date to 2011

MSA	Percent Chg. SR	Percent Chg. CSW	Difference
Atlanta	-22.7	-29.7	7.0
Boston	-16.5	-25.1	8.6
Charlotte	-6.1	-22.9	16.8
Chicago	-20.4	-37.5	17.1
Cleveland	-19.5	-27.7	8.3
Dallas	-2.6	-15.9	13.4
Denver	-9.7	-18.5	8.8
Detroit	-39.2	-51.4	12.2
Las Vegas	-52.8	-63.1	10.2
Los Angeles	-18.0	-43.9	25.9
Miami	-36.3	-54.7	18.5
Minneapolis	-23.8	-40.7	16.8
New York	-14.1	-30.3	16.2
Phoenix	-42.7	-60.0	17.4
Portland	-21.4	-32.9	11.5
San Diego	-25.6	-44.5	18.9
San Francisco	-18.5	-44.2	25.7
Seattle	-20.0	-33.6	13.6
Tampa	-37.0	-51.2	14.1
Washington, DC	-21.3	-34.0	12.7
Median			13.9

Table 4: Estimates of Parameters

Nonlinear Least Squares Estimates

	Parameter	Estimate	Standard Error
(1)	ρ	1.0473	0.0382
(2)	κ	0.5281	0.0164

Maximum Likelihood Estimates

	Parameter	Estimate	Standard Error
(3)	ρ	0.9924	0.0353
(4)	κ	0.5537	0.0122
(5)	σ_u	0.0454	0.0030

Figure 1: Comparison of CSW to Average of Self Reports (SR)

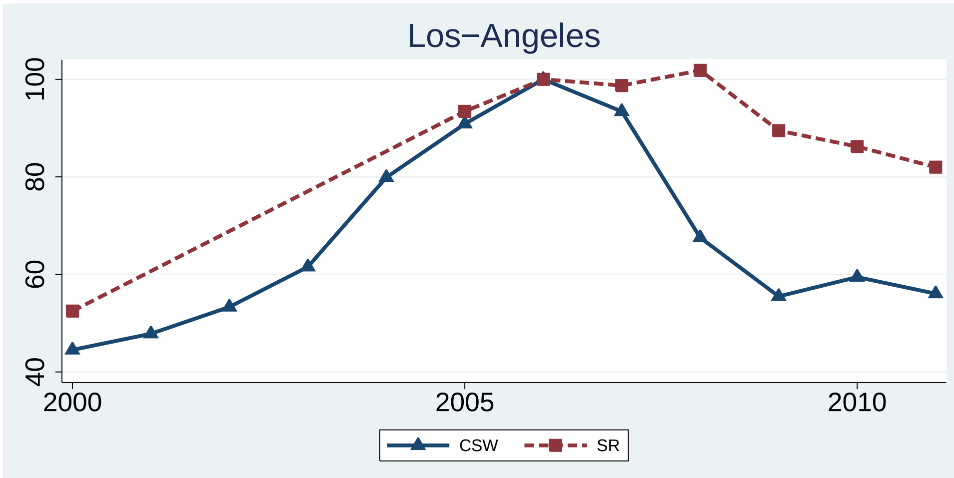
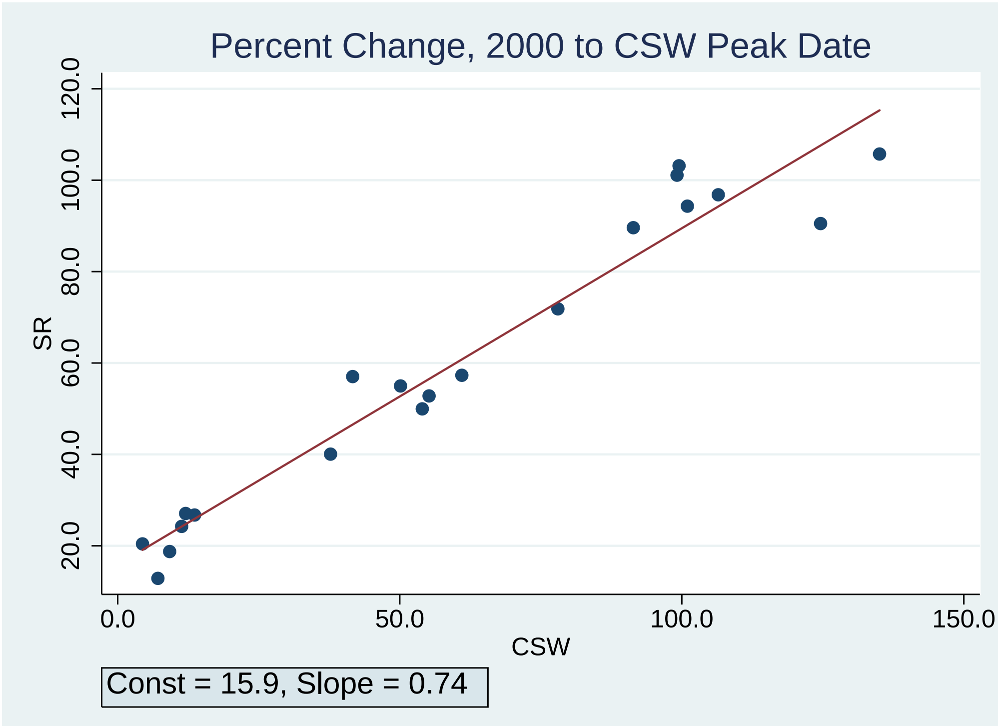


Figure 2: Comparison of Real Growth, CSW and Average of Self Reports (SR)



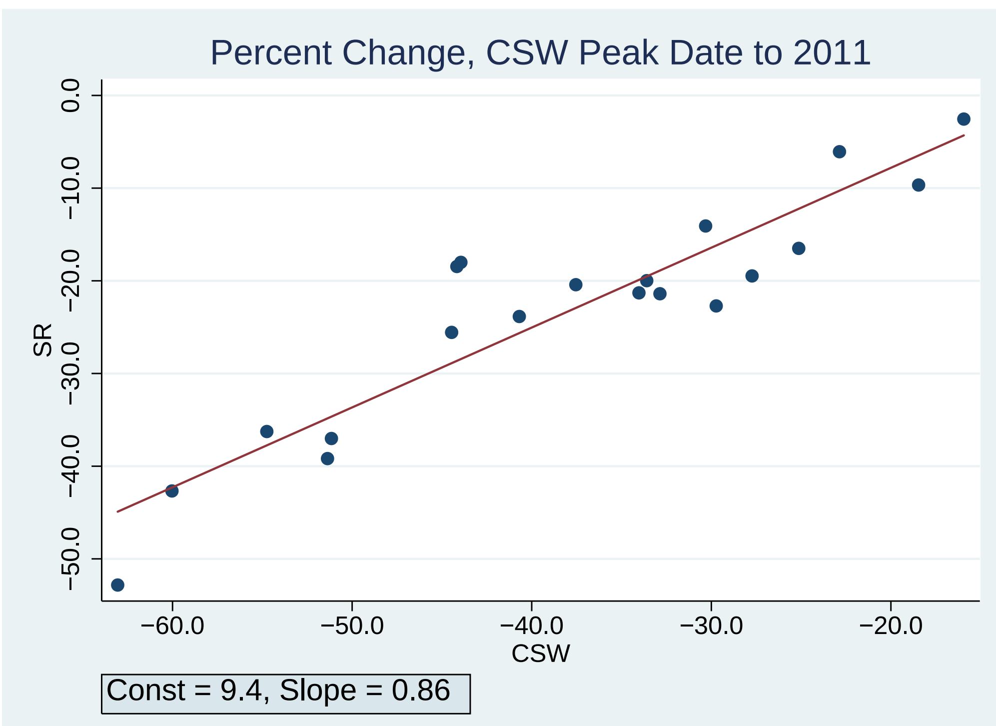
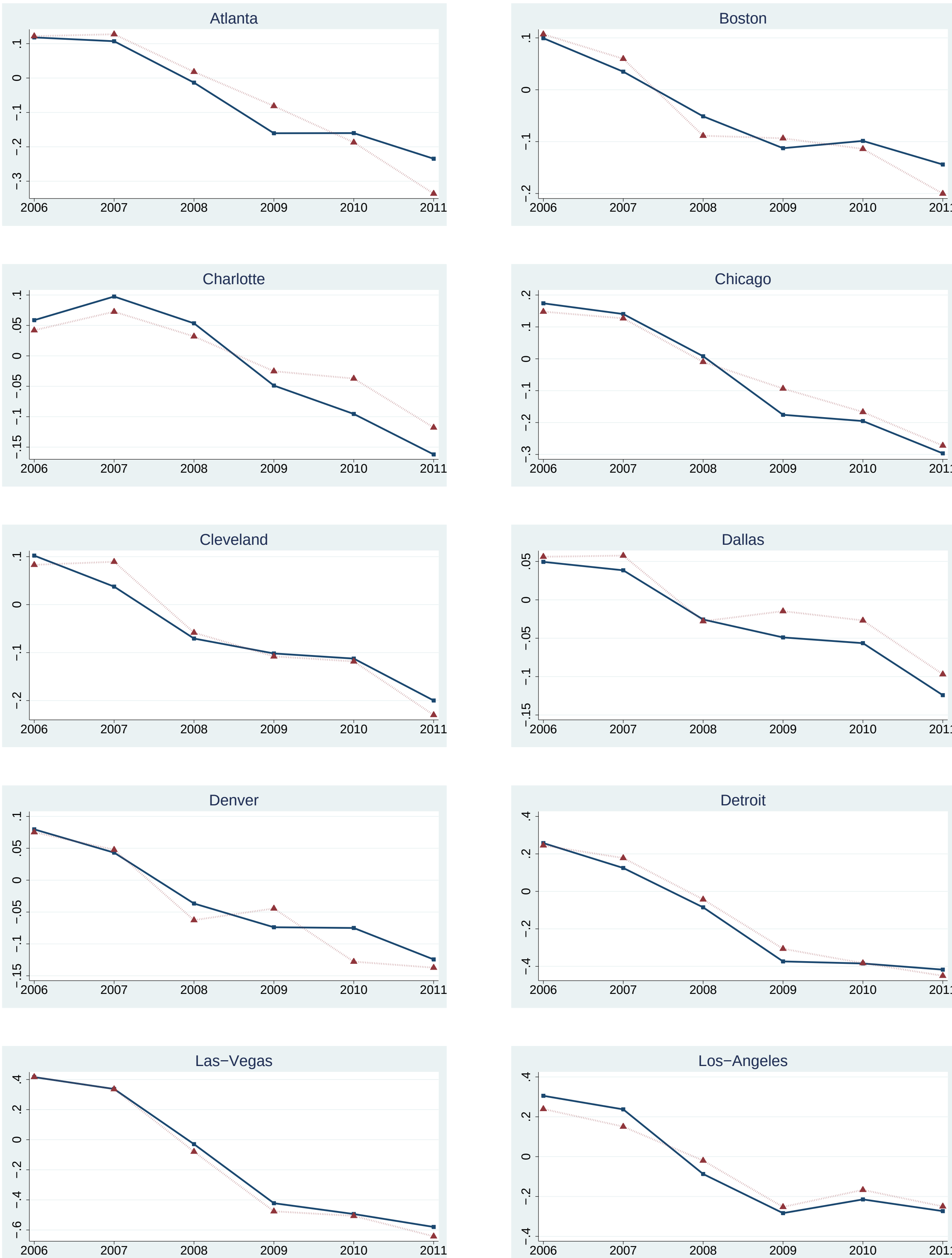


Figure 3: Comparison of $\mathcal{H}_{st}$ (square) and estimate of H_{smt} (triangle)



Comparison of $\mathcal{H}_{st}$ (square) and estimate of H_{smt} (triangle)

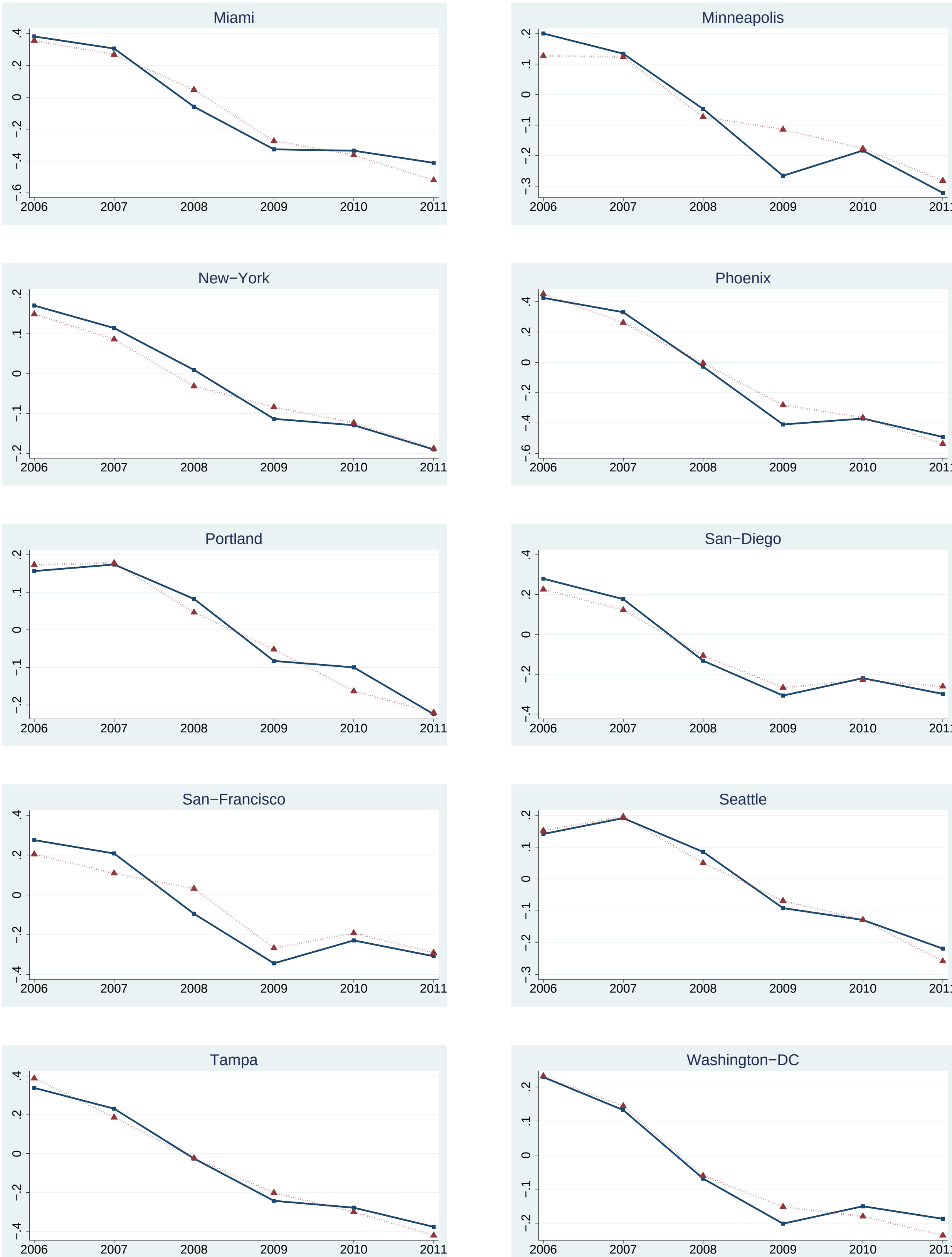
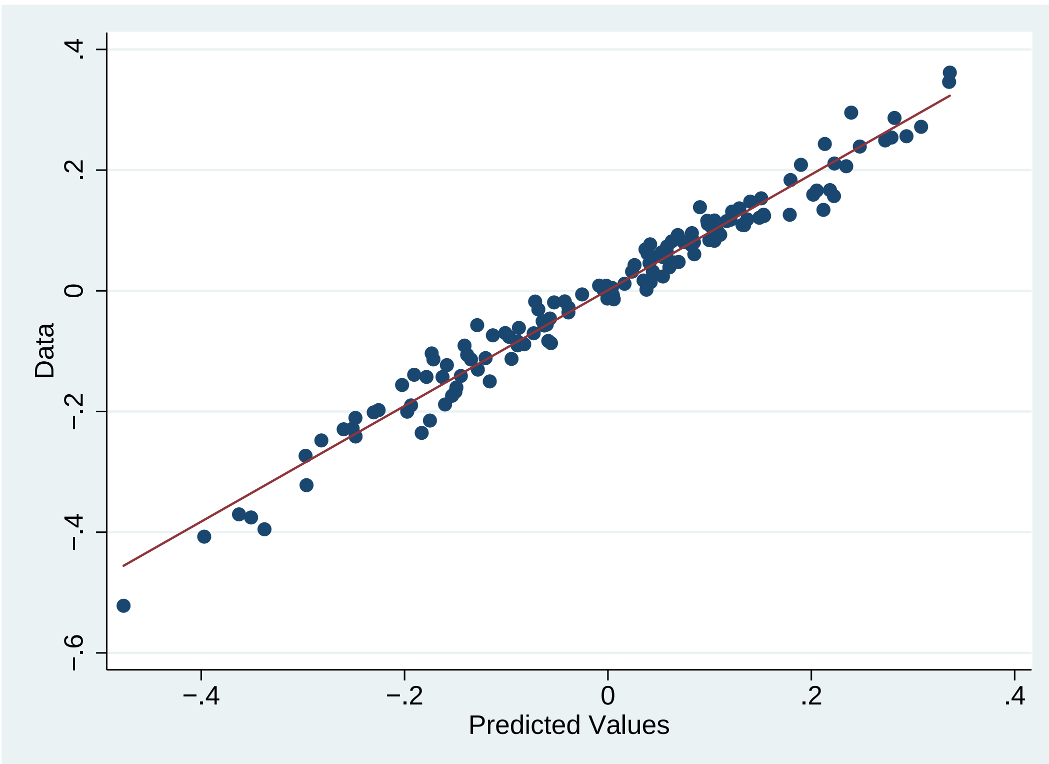


Figure 4: Comparison of data on H_{bt} to predicted values, 2005-2011



$$\underset{}{\text{Data}} = \underset{(0.002)}{0.001} + \underset{(0.014)}{0.959} * \text{Predicted}$$