NBER WORKING PAPER SERIES

ESTIMATING LOAN-TO-VALUE AND FORECLOSURE BEHAVIOR

Arthur Korteweg
Morten Sorensen

Working Paper 17882
http://www.nber.org/papers/w17882

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2012

We are grateful to Chris Mayer and the Paul Milstein Center for Real Estate at Columbia Business School for generous access to data and research assistance. We thank Chris Downing, Dasol Kim, Monika Piazzesi, Stijn van Nieuwerburgh, Karen Pence, Chester Spatt, Richard Stanton, NancyWallace, along with seminar participants at the Berkeley Capital Markets Real Estate seminar, Columbia Business School, Duke, Haas School of Business, NBER Summer Institute (2011), Santa Fe Summer Real Estate Symposium (2011), Stanford Graduate School of Business, Rice, SIFR, Stockholm EFA (2011), Utah, and UT Dallas for helpful comments and feedback. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimating Loan-to-Value and Foreclosure Behavior
Arthur Korteweg and Morten Sorensen
NBER Working Paper No. 17882
March 2012
JEL No. C11,C23,C24,C43,R21,R3

ABSTRACT

We develop and estimate a unified model of house prices, loan-to-value ratios (LTVs), and trade and foreclosure behavior. House prices are only observed for traded properties, and trades are endogenous, creating sample-selection problems for traditional estimators. We develop a Bayesian filtering procedure to recover the price path for each individual property and produce selection-corrected estimates of historical LTVs and foreclosure behavior, both showing large unprecedented changes since 2007. Our model reduces the index revision problem by nearly half, and has applications in economics and finance (e.g., pricing mortgage-backed securities).

Arthur Korteweg
Graduate School of Business
Stanford University
518 Memorial Way
Stanford, CA 94305-5015
korteweg@stanford.edu

Morten Sorensen
Columbia University
Columbia Business School
Uris Hall 802
3022 Broadway
New York, NY 10027
and NBER
ms3814@columbia.edu

Real estate price indices, loan-to-value ratios (LTVs), and trade and foreclosure behavior are important measures of economic activity. Yet, these measures are difficult to estimate, and the literature contains surprisingly little formal analysis of the resulting inference problems. To our knowledge, this paper contains the first formal econometric analysis of the problem of estimating LTV distributions. Additionally, it addresses open issues in the closely related inference problems for price indices and trade and foreclosure behavior.

The difficulty of estimating LTV distributions arises because a property's price (the denominator in the LTV ratio) is only observed when the property trades, and trades are endogenous. Two existing approaches to this problem are surveys and imputed prices, both with serious limitations. LTV estimates from surveys (e.g., the Survey of Consumer Finances (SCF) and the American Housing Survey (AHS)) are limited by the scope and timing of the surveys. More importantly, home owners may neither be fully informed nor entirely objective when valuing their home hypothetically for a survey, creating methodological concerns and substantial disparities between overlapping surveys (e.g., Melzer, 2011). LTV estimates from imputed prices (e.g., from CoreLogic and Zillow) start from the price or LTV that is observed when a property is traded or refinanced and then impute an updated price using a local price index or hedonic pricing model (see Bhutta et al. (2010), Landier et al. (2011), Laufer (2011), and Okah and Orr (2010) among others). Imputing prices implicitly assumes that properties appreciate exactly as predicted by the model, i.e., that there is no idiosyncratic volatility. With idiosyncratic volatility, this approach underestimates the dispersion in prices and LTVs, resulting in downward biased estimates of the fraction of underwater properties. To illustrate, if a local price index shows an average appreciation of 5%, it matters whether this represents exactly 5% appreciation for each individual property, or whether it represents, say, a 15% increase for one half, and a 5% decline for the other half. By assuming the former, the imputation approach mechanically underestimates the dispersion. Additionally, idiosyncratic volatility introduces a dynamic sample-selection bias, as described below.

We present a new econometric model of house prices, LTV ratios, and trade and foreclosure behavior, and we estimate this model using data from Alameda County, California, spanning 1988 to 2008. Our results suggest that selection biases are substantial. We es-

timate an idiosyncratic annual volatility in individual home prices of 28%. Towards the end of our sample period, a naïve imputation of property prices using a repeat-sales index leads to an estimate of the fraction of underwater properties of 20%. Accounting purely for the idiosyncratic volatility and the resulting increase in the dispersion of prices (but not dynamic selection) increases this estimate to 30%. Different specifications of the dynamic selection effect result in final estimates of the fraction of underwater properties in the 28%–37% range.

For an external validation of our estimates, we investigate the index revision problem. This is a well-known sample-selection problem in which the current estimates of the most recent index value are systematically revised (typically downwards) when subsequent data become available. In our baseline specification, this revision problem is reduced by about 45% (average long-term downward revisions are reduced from 13% to 7%), confirming that our model captures a substantial amount of the dynamic selection of traded and foreclosed properties.

The difficulties of estimating LTVs lead to other inference problems. Trade and foreclosure decisions are endogenous and depend on current prices and LTVs (e.g., Case, Pollakowski, and Wachter, 1997). This creates a selection problem when estimating prices and price indices.[1] Correcting such price estimates requires a selection model, such as a probit or hazard model of trade and foreclosure behavior. This selection model, however, requires current LTVs or prices as explanatory variables. We resolve this circularity by jointly estimating the price index and the trade and foreclosure model.

Our approach extends the standard repeat-sales model by Bailey, Muth, and Nourse (1963) and Case and Shiller (1987) by modeling the entire price path for each individual

[1]This problem is aptly described by Calnea (Calnea Analytics (2010), p. 14), who calculates the United Kingdom's land-registry house-price index using a repeat-sales regression (RSR): "The RSR index is naturally more reflective of properties that transact more frequently. In so far as a differential in price appreciation exists between properties based on the relative frequency of transactions, the RSR measure will be naturally weighted towards the more frequently transacting subset of properties. There are a variety of reasons why the holding duration of properties might be unevenly distributed. The increase in transaction costs for more expensive properties due to stamp duty may result in a decreased turnover of more expensive homes. "Life-cycle" theories on property holding period posit that less expensive properties are traded more frequently—when people move up the property "ladder" they tend to move home less often. In addition the Buy-to-Let market is more active in the lower price brackets. Policy-makers need to be aware of the price appreciation differentials between sub markets, especially when there is systematic variation in the frequency of transactions between these sub markets."

property, even when it is not traded, and by explicitly modeling trade and foreclosure decisions for each property at each point in time. Simultaneously estimating prices, LTVs, and trade and foreclosure behavior has several advantages. First, econometrically, the trade process is a dynamic extension of the selection equation in the standard sample-selection model (Heckman, 1979, 1990), and simultaneously estimating the price and trade processes corrects for selection bias. (For this reason we sometimes refer to the trade process as the *selection* or *trade-selection process*.) Second, the parameters in the trade process reflect trading behavior. Separating regular sales from foreclosure sales and allowing coefficients to vary over time, we find large and unprecedented changes in recent behavior. Third, the approach estimates the full cross-sectional and time-series distributions of the LTVs. Finally, our procedure consistently incorporates price information revealed by untraded properties. To illustrate, when appreciating properties are more likely to trade, a period with few trades suggests declining prices. (Conversely, when *depreciating* properties are more likely to trade, possibly in foreclosure sales, periods with few trades suggest appreciating prices; our model captures both cases.) Hence, trading volume is informative about prices and should be included in the estimator. This is fundamentally different from the standard Heckman (1979) cross-sectional sample-selection model. In this standard model, observations with unobserved outcomes are only informative about the first stage, not the second one.[2] In our dynamic extension, prices are serially correlated, and the fact that a property is not traded is informative about its current price, which is informative about its previous and subsequent prices due to the serial correlation. Hence, observations with unobserved outcomes must be included in the second stage as well.[3]

Our Bayesian estimator has several advantages as well.[4] First, modeling price dynamics for individual properties is numerically intensive, rendering standard maximum likeli-

[2]In the standard selection model, unobserved outcomes are independent and can be integrated out of the likelihood function for the second stage, conditional on the estimates of the first stage. With serial correlation, unobserved outcomes are no longer independent and cannot be integrated out.

[3]Some studies construct corrected real-estate indices using a standard Heckman selection model (e.g., Jud and Seaks, 1994; Munneke and Slade, 2000; and Hwang and Quigley, 2004). The standard Heckman model, however, does not allow serial correlations in unobserved outcomes, and the price information revealed by untraded properties is not incorporated, raising concerns about whether these model are correctly specified and whether the resulting estimators are consistent.

[4]Our procedure is substantially different from the Bayesian estimators of the repeat-sales model described and compared by Goetzmann (1992), which do not exploit these advantages to the same extent.

hood (ML) estimation infeasible. Bayesian estimation, however, remains computationally feasible with robust convergence properties, as explained below. Second, the estimation procedure is a numerical Monte Carlo procedure that simulates unobserved prices as part of the estimation. This makes it straightforward to construct posterior distributions of the estimated parameters and non-linear transformations of these, such as price indices, LTVs, and foreclosure and trade intensities. In contrast, standard statistical inference (e.g., calculating standard errors) is difficult for such transformations, especially involving variance parameters, which are not asymptotically normal. Finally, the Bayesian estimator produces accurate small-sample inference. Although our sample contains around 70,000 properties, the number of estimated parameters and prices is also large, and it is difficult to assess whether asymptotic approximations would be appropriate.

As a final comment, our model extends the empirical finance literature about illiquid assets with unobserved prices, such as private equity and venture capital investments in privately held companies (Cochrane, 2005; and Korteweg and Sorensen, 2010). The empirical issues are closely related,[5] and our approach may be useful for understanding prices and trading behavior of other illiquid or asynchronously-traded assets (e.g., corporate bonds, small-cap stocks, and index arbitrage). More generally, the empirical approach exploits recent advances in Bayesian computational procedures—specifically, Markov Chain Monte Carlo (MCMC), Gibbs Sampling, and forward filtering backwards sampling (FFBS)—that permit estimation of models with infrequently observed behavior of individual agents, which may be useful for other applications.

In Section I, we present the empirical model and discuss the estimation procedure and identification. Section II presents the data. In Section III, we compare price indices estimated with and without correcting for sample selection. Section IV presents estimates of the LTV distribution. In Section V, we analyze trade and foreclosure behavior. Section VI contains concluding remarks. Details about the estimation procedure are in the Appendix.

[5]Bloomberg (August 17, 2011) reported that there is "a record number of private-equity firms raising real estate funds," suggesting deeper economic relationships between real estate and private equity.

I Model of Prices and Trades

To fix ideas and notation for our empirical model, it is useful to derive the discrete-time price dynamics from continuous-time fundamentals. Let $\mu(t)$ be a common exogenous determinant of price appreciation (a price index), and let the price of property i follow the Brownian motion:

$$\frac{dP_i(t)}{P_i(t)} = \mu(t)dt + \sigma_i dB_i(t). \tag{1}$$

Define the log-price $p_i(t) = \ln(P_i(t))$ and let $\delta_i(t) = \int_{t-1}^{t} \mu(\tau)d\tau - \frac{1}{2}\sigma_i^2$. Using Ito's lemma, the change in the log-price from time t to t' is:

$$p_i(t') = p_i(t) + \left[\sum_{\tau=t+1}^{t'} \delta_i(\tau)\right] + \varepsilon_i(t,t'), \tag{2}$$

with $\varepsilon_i(t,t') \sim N\left(0,(t'-t)\sigma_i^2\right)$.

The standard repeat-sales regression (RSR) is estimated from this equation using properties that trade (at least) twice by regressing the change in the observed log prices on indicator variables for the intermediate periods between the trades, represented by the $\delta_i(\tau)$ terms in the equation. This regression is typically implemented using generalized least squares (GLS) to correct for heteroscedasticity by weighing each observation by the inverse of the square root of the time between trades. With a sufficient number of partially overlapping trades, all $\delta(t)$ coefficients are identified (assuming no heterogeneity in σ_i). The estimated $\delta(t)$ coefficients are consistent when ε is independent of the indicator variables (i.e., when the expected value of ε is independent of whether or not the property trades). This independence fails, however, when the decision to trade is not independent of the price appreciation, $\varepsilon_i(t)$. In this case, the error term is correlated with the indicator variables, creating the sample-selection problem and potentially biasing the estimated coefficients.

Setting $t = t' - 1$, the one-period transition equation is:

$$p_i(t) = p_i(t-1) + \delta(t) + \varepsilon_i(t), \tag{3}$$

with $\varepsilon_i(t) \sim N(0, \sigma_i^2)$. Since most prices are unobserved, this equation cannot be estimated directly. Treating the price as an unobserved state variable, however, our estimation procedure uses this equation to filter out the unobserved prices between trades. Following the finance terminology, we also refer to $\delta(t)$ as *systematic variation* and $\varepsilon_i(t)$ as *idiosyncratic volatility*.[6]

The estimated $\delta(t)$ coefficients can be transformed into price indices. Normalizing by the price level at time t_0, it is natural to define an index as the population average of current prices relative to time-t_0 prices (in levels, not logs) as:

$$I(t) = E\left[\frac{P_i(t)}{P_i(t_0)}\right] = E\left[\exp\left(\left[\sum_{\tau=t_0+1}^{t} \delta(\tau)\right] + \varepsilon_i(t_0,t)\right)\right] = \prod_{\tau=t_0+1}^{t} \exp\left(\delta(\tau) + \tfrac{1}{2}\sigma^2\right). \quad (4)$$

The one-period change in this index is $I(t)/I(t-1) = \exp\left(\delta(t) + \frac{1}{2}\sigma^2\right)$. Note that typical repeat-sales indices, such as those from S&P/Case-Shiller, CoreLogic, and FHFA (formerly OFHEO), are defined without the $\frac{1}{2}\sigma^2$ adjustment term. Below, we compare indices with and without this adjustment. Goetzmann (1992) discusses this adjustment in more detail, and denotes indices with and without it as *arithmetic* and *geometric* indices, respectively.

A Trade and Foreclosure Processes

To model sales, it is convenient to define the latent discrete time process $w_i(t)$, such that property i trades between time $t-1$ and t, and hence $p_i(t)$ is observed, when

$$w_i(t) \geq 0. \quad (5)$$

This trade process is parametrized as:

$$w_i(t) = W_i'(t)\alpha_0 + p_i(t)\alpha_p + \eta_i(t), \quad (6)$$

[6] One can imagine estimating more flexible price processes, such as auto-regressive processes or processes that depend on property characteristics or whether properties trade in foreclosure sales. This raises issues about the identification of the model, left for future research. A main advantage of the simple specification used here is that it is equivalent to the standard repeat-sales model, making the results directly comparable.

where $\eta_i(t)$ is $i.i.d.$ $N(0,1)$. This specification is equivalent to a binary probit model, and the parameters are only identified up to scale. Without loss of generality, the scale is normalized by fixing the error term's variance to one. The term $p_i(t)$ is the log price. Property characteristics and prevailing mortgage rates (and a constant term) are in $W_i(t)$, along with the log loan amount with a coefficient fixed to -α_p, implying that α_p is the negative of the coefficient on log LTV.

The resulting model is a dynamic extension of the standard cross-sectional sample-selection model from Heckman (1979, 1990) with the trade process as the selection process. Under standard conditions, jointly estimating these processes corrects for the selection bias that may arise when the price dynamics of properties with observed prices are not representative of the price dynamics in the population overall. Note that the trade process depends on the contemporaneous LTV ratio. This is a natural specification, yet it has been difficult to implement in existing studies, because contemporaneous LTVs are unobserved. Our model circumvents this problem by jointly estimating the latent price and trade processes.

In our data, normal sales are distinguished from foreclosure sales, and one might suspect, as we find empirically, that these sales follow different processes. Appreciating properties are more likely to trade in normal sales, whereas depreciating properties are more likely to end in foreclosures. Hence, we also estimate specifications with separate trade and foreclosure processes. In these specifications, we include an additional foreclosure process such that a foreclosure sale (but not necessarily a normal sale) occurs between time $t-1$ and t when

$$z_i(t) \geq 0. \tag{7}$$

This process is parametrized as:

$$z_i(t) = Z_i'(t)\gamma_0 + p_i(t)\gamma_p + \xi_i(t). \tag{8}$$

As above, ξ is $i.i.d.$ $N(0,1)$, and $Z_i(t)$ contains observed variables that affect the probability of foreclosure sales, including log LTV, property characteristics, and mortgage rates. In our empirical specifications, we always have $Z = W$.

B Empirical Implementation

To summarize, the baseline model specifies the price and trade processes:

$$p_i(t) = p_i(t-1) + \delta(t) + \varepsilon_i(t), \tag{9}$$

$$w_i(t) = W_i'(t)\alpha_0 + p_i(t)\alpha_p + \eta_i(t), \tag{10}$$

$$w_i(t) \geq 0 \Leftrightarrow p_i(t) \text{ is observed.} \tag{11}$$

The price process, $p_i(t)$, is mostly unobserved, except when $w_i(t) \geq 0$. The trade (or selection) process, $w_i(t)$, is entirely unobserved. The vector $W_i(t)$ is observed data. The error terms are *i.i.d.* with $\varepsilon_i(t) \sim N(0, \sigma^2)$ and $\eta_i(t) \sim N(0,1)$. The estimated parameters of interest are α, σ^2, and $\delta(t)$.[7]

This model defines a likelihood function. ML estimation is complicated, however, by the large number of latent variables, since evaluating the likelihood requires integrating over them jointly. We specify the model at the quarterly level over a 20-year period, resulting in 160 ($= 2 \times 4 \times 20$) latent variables per property. With around 70,000 properties, each evaluation of the likelihood requires numerically evaluating 70,000 160-dimensional integrals, rendering ML estimation numerically intractable. As a feasible alternative, we develop a Bayesian procedure, using an MCMC method known as Gibbs sampling, to substantially reduce the computational burden. We provide an overview of this procedure below, and more details are in the Appendix (see also Korteweg and Sorensen, 2010).

The model is constructed such that its variables can be divided into three blocks. The first block contains the parameters, α, σ^2, and $\delta(t)$; the second one contains the variables in the trade processes, $w_i(t)$; and the third block contains the price processes, $p_i(t)$. The Gibbs sampler simulates the (augmented) posterior distribution by iteratively drawing the variables in each block conditional on the previous draw of the variables in the other blocks (for a formal treatment see Geman and Geman (1984), Tanner and Wong (1987), Gelfand

[7] Although not pursued here, these assumptions can be relaxed somewhat. It is possible to estimate specifications with property-specific coefficients using hierarchical priors, and the normality assumptions generalize to mixtures of normals without losing tractability of the Bayesian procedure. The log-price must enter the trade process linearly, however, to maintain a linear Kalman filter, since generalizing the log-linear specification introduces substantial numerical complexity. See Korteweg and Sorensen (2010) for details.

and Smith (1990), Johannes and Polson (2006), and Korteweg (2012)).

In the first block, conditional on the previous draw of the price and trade processes, the parameters α, σ^2, and $\delta(t)$ are given by the linear equations (9) and (10), and they are drawn from a standard Bayesian linear regression.[8] Drawing the trade process in the second block is similarly straightforward. Conditional on the parameters and prices, the distribution of $w_i(t)$ follows a truncated normal distribution that is constrained to be negative when there is no trade and positive when a trade is observed, as specified by equation (11). These first two blocks are analogous to Bayesian estimation of probit models (see Albert and Chib, 1993). The key to the model is the third block, where the entire path of unobserved prices is drawn using the FFBS procedure (Carter and Kohn, 1994; and Fruhwirth-Schnatter, 1994). Conditional on the parameters and trade process, the price process can be viewed as being defined by a linear state space or Kalman filter. Under this view, $p_i(t)$ is the unobserved state variable; the transition rule is the one-period price equation (9); the index $\delta(t)$ is an "observed" control acting on the state; and, conditional on $w_i(t)$, the trade process is an observation equation, providing noisy "observations" of the state. Given this setup, the FFBS procedure draws from the conditional posterior distribution of the entire price path, as required in the third block of the Gibbs sampler.[9]

This Gibbs sampling procedure iteratively draws from the joint posterior distribution of the parameters and the individual price paths for the properties. Using these draws, the posterior distributions of the price index and LTV distributions are straightforward to construct. For the price index, fixing t, in each iteration, we calculate $I(t) = \prod_{t_0+1}^{t} \exp\left(\delta(\tau) + \frac{1}{2}\sigma^2\right)$ using the current draws of $\delta(t)$ and σ^2. Across iterations, the resulting distribution is the posterior distribution of $I(t)$.[10] Repeating this calculation for all t generates the posterior distribution of the entire index.

The construction is slightly more complicated conceptually for the LTV distribution,

[8] Technically, this is implemented as draws from several blocks, as explained in detail in the Appendix.

[9] While it is convenient to describe the blocks in this order, the actual sampling procedure has better numerical properties when starting with the prices in the third block. By setting the initial value of $\alpha_p = 0$, the first iteration can draw the prices without specifying initial values of the trade process, which speeds up convergence.

[10] Note that it would be difficult to perform standard classical asymptotic inference on this index, such as calculating its standard error, since $I(t)$ is a nonlinear function of the estimated parameters and the asymptotic distribution of σ^2 is not normal.

since its posterior is a distribution of distributions (or, more precisely, a time series of the distribution of distributions). In each iteration, the $LTV_i(t) = Principal_i(t)/P_i(t)$ is calculated using the outstanding principal (described below) and the current draws of the price processes. Collecting these values across properties produces one draw from the posterior distribution of the cross-sectional distribution of LTVs. Collecting these cross-sectional distributions across iterations and time produces the time series of their posterior distributions. From these distributions it is straightforward to calculate, for example, the time series of the estimated fraction of underwater properties, i.e., properties with LTVs exceeding one.

C Identification

Heckman (1990) shows that semi-parametric identification of the standard selection model requires exogenous or predetermined variation in the selection equation. We include the time since the previous trade (*Time*) as such variation. Following the logic of the standard model, there are two requirements for *Time* to be a valid. First, the exclusion restriction requires that *Time* is independent of the error term in the price process. When the price process follows a martingale, which is a common assumption for price processes, this exclusion restriction holds mechanically. Second, *Time* must be directly related to the probability of a sale. Due to transaction costs, it is reasonable to think that a new owner does not intend to resell a property immediately after buying it. Hence, immediately after a trade, the trade intensity declines and then gradually increases independently of the idiosyncratic term in the price process. This behavior is consistent with the well-known phenomenon of "seasoning" of mortgage-backed securities, where new loans prepay slower than older loans, and our empirical results confirm this pattern. With these two requirements, *Time* provides valid exogenous variation for the identification of the model. Note, however, that the empirical results are very similar for specifications with and without *Time*, suggesting that the model is reasonably well identified from distributional assumptions alone.

For a formal identification argument, consider properties trading at time t_0, some of which trade again one period later, at time $t_0 + 1$. This simple case is equivalent to the standard Heckman model, and identification follows from Heckman's (1990) identification

argument, which requires a variable, $s(t_0+1)$, that is independent of $\varepsilon(t_0+1)$, and that the probability of a trade at time t_0+1 is a non-degenerate function of $s(t_0+1)$. Our case is more complex, though, since the predetermined source of variation is the time since the previous trade, and we cannot only consider trades one period apart. For our case, compare properties trading at time t_0 and again at time $t_0+\tau$ to properties trading repeatedly at times $t_0, t_0+1, t_0+2, ..., t_0+\tau$. The average appreciation for the first properties is $\sum_{t=t_0+1}^{t_0+\tau} E\left[\delta(t)|\, s=\tau\right] = E\left[\left(\sum_{t=t_0+1}^{t_0+\tau}\delta(t)\right)|\, s=\tau\right]$. The average appreciation for the repeatedly trading properties is $\sum_{t=t_0+1}^{t_0+\tau} E\left[\delta(t)|\, s=1\right] = E\left[\left(\sum_{t=t_0+1}^{t_0+\tau}\delta(t)\right)|\, s=1\right]$. Hence, comparing the appreciation of properties trading rarely to properties trading frequently over the same period of time provides observations of the same sum of $\delta(t)$ under various degrees of selection. Varying the predetermined variable, s, leads to different amounts of selection, and the standard identification argument applies.

As a more intuitive illustration, compare property A, trading at times 0 and 2, to property B, trading at times 0, 1, and 2 (each time period may span several years). The total appreciation from time 0 to 2 is observed for both properties and denoted δ_A and δ_B.[11] Assume that trading one period apart (property B) is unusual and has a lower probability than trading two periods apart (property A). These probabilities are observed (in the standard Heckman model, they are estimated in the "first stage"). Given that property B traded in an unusual way, it may have experienced a price shock that compensated for the low probability. If $\delta_B > \delta_A$, property B traded more frequently because it experienced a positive shock, and we infer that the coefficient α_p in the trade-selection equation is positive. Conversely, if $\delta_B < \delta_A$, we infer $\alpha_p < 0$. The magnitude of α_p can be derived from the difference $\delta_B - \delta_A$ relative to the difference in the probabilities. Finally, we might observe $\delta_B = \delta_A$ and infer that trades are independent of price and $\alpha_p = 0$. Given the identification of α_p, the identification of the remaining parameters is straightforward.

[11] For a more formal identification argument, each of these properties represents a sufficiently large number of individually trading properties that the law of large numbers applies and individual error terms can be replaced by their conditional expectations. Specifically, δ_A and δ_B are the expected appreciations conditional on the different trading patterns.

II Data

We use data from transactions of single-family residences in Alameda County, California. Alameda County is located in the San Francisco East Bay Area and includes, amongst others, the cities of Oakland and Berkeley. The data are from CoreLogic (formerly First American), obtained through the Paul Milstein Center for Real Estate at Columbia Business School, and they include all transactions in the 20-year period from 1988:Q1 to 2008:Q3. Unfortunately, the center's agreement with CoreLogic expired in 2008, and we have been unable to extend the sample beyond this period. The data contains sales dates and prices, mortgage amounts, and refinancing information obtained from the deeds records. In addition, tax records contain information about each property's characteristics, such as size, number of bedrooms, single- or multi-family residence, etc.

We restrict the sample to properties that satisfy the following criteria: the property is a single-family residence; it trades at least twice during our sample period, and both sales are full sales (not partial sales); the tax records have no missing property characteristics; the property's characteristics have not changed between the two sales; it has no more than 10 bedrooms, 5 bathrooms, and 3 stories; and it is located on less than 5 acres with less than 10,000 square feet of living space. The resulting sample contains 164,824 transactions of 68,700 properties. Table I presents summary statistics, and Figure 1 shows the time series of normal trades and foreclosure sales within the sample.

While the data contain initial mortgage amounts and subsequent refinancings, they contain no information about amortization schedules. We estimate an amortization schedule assuming a 30-year fixed-rate mortgage at the prevailing rate at the time of the transaction. Whenever possible we update the outstanding mortgage amounts using refinancing information.[12] Using these calculated outstanding loan principals, we calculate the LTV, with the value in the denominator given by the price process specified in the model.

This calculation fails for transactions of properties without mortgages. These properties have LTVs equal to zero, leaving log LTV undefined. We set their log LTV to -3, corresponding to an LTV ratio of 5%. These cases are infrequent, and the particular number has

[12] Refinancings typically involve an appraisal of the property's value. These appraisal values are *not* used for estimating the price process.

a negligible effect on our estimates, as does excluding them altogether. We do not truncate the upper tail of the LTV distribution, but in unreported estimates, we find that doing so has a negligible impact on the estimates. The top plot in Figure 3 shows the distribution of the buyers' LTV ratios at the time of the transactions. The bottom plot shows the sellers' distribution. Both LTVs are observed at the time of a transaction, and these plots only include those observed ratios, not those estimated by the model.

Foreclosure sales are identified in the data, typically separating the initial seizure of the property from the subsequent sale of the repossessed property by the bank. We define the latter of those two transactions as the foreclosure sale. The former transaction, when a property is seized by the lender, is not an arms-length transaction and does not have a well-defined price or mortgage, and it is not included in the estimation.

Mortgage rates are collected from FRED, a database maintained by the Federal Reserve in St. Louis. We use the change in the prevailing mortgage rate since the transaction in the specifications. When the mortgage rate declines, it becomes more attractive for owners to sell or refinance their loans.

III Price Dynamics

We first compare the price dynamics without the trade process and selection correction. Figure 3 plots indices estimated using the standard GLS procedure and our MCMC procedure (without selection correction, i.e., fixing $\alpha_p = 0$), along with the S&P/Case-Shiller Home Price Index for the San Francisco metro area.[13] Arithmetic indices include the $\frac{1}{2}\sigma^2$ adjustment (the annualized GLS estimate of σ is 0.2811). Geometric indices are calculated without this adjustment. All indices are normalized to 100 in 2000:Q1.

Apart from the large differences between arithmetic and geometric indices, the indices are very similar. The S&P/Case-Shiller Index is similar to the geometric indices calculated here, although Case-Shiller is about ten points lower at the peak. This ten-point difference probably arises because the S&P/Case-Shiller Index includes transactions from the entire San Francisco metropolitan statistical area (MSA), comprising the counties of Alameda,

[13] Downloaded from the S&P/Case-Shiller website on February 25, 2010.

Contra Costa, Marin, San Francisco, and San Mateo; our data cover only Alameda County, which is around one-third of the MSA by population. Additionally, unlike our indices, the S&P/Case-Shiller Index is smoothed and value-weighted.

In Figure 3, the indices from the GLS and MCMC estimators appear virtually identical. Econometrically, the GLS estimator is defined in terms of moment conditions and the assumption that the variance of the error term increases linearly with the time between trades. The Bayesian MCMC estimator also imposes distributional assumptions on the error terms and priors. The similarity between the resulting indices suggests that, at least absent the trade process and selection correction, the MCMC estimator's distributional assumptions are reasonable.

Figure 4 plots indices corrected for selection using the trade process. Models A to F refer to the different specifications of this process, given in Table II. The baseline model is the MCMC estimator without selection correction from Figure 3. Unlike the corrected indices, the uncorrected index from the baseline model shows a small increase by the end of the sample period. Figure 5 shows similar indices estimated from specifications with separate processes for normal trades and foreclosure sales. This leads to slightly larger effects of the selection correction. Most pessimistic is Model A, producing a final index around 140. Model F is most optimistic, ending with the index at 170. In contrast, the baseline index ends at a level around 155.

Indices with and without selection corrections are broadly similar, suggesting that long-term biases from sample selection in price indices are, perhaps, a smaller concern than previously thought (e.g., by Calnea Analytics, see footnote 2). Comparing the two indices, the quarter-by-quarter changes can be quite different, but over the sample period, the levels never diverge substantially. Intuitively, while more rapidly appreciating properties trade during the price run-up, eventually the more slowly appreciating properties also trade. When they trade, they bring the index back to the underlying population average, and while the standard index may show temporary biases due to selection, as long as all properties eventually trade, the levels of the standard and selection-corrected indices are unlikely to diverge substantially.

A Index Revisions

In the short term, selection bias leads to the well-known index revision problem (e.g., Clapham et al. (2006)). The index revision problem is the sample-selection problem that arises when initial transaction data are selected and updated with new trades, resulting in revisions of the index. In other words, the sample selection problem is important in the short run, but disappears in the long run. Systematic revisions of the index value for a given period as more data arrives have, for example, made it difficult to construct financial derivatives based on the index value.

To test the external validity of our model, we investigate whether our selection-corrected index is less susceptible to the index revision problem than the standard repeat-sales index. Since the index revision problem is a sample-selection problem, this provides a natural test of the model.[14] Figure 6 shows index estimates where the sample is extended incrementally by one quarter at a time, and the entire index is re-estimated from the extended sample (for expositional clarity, the plot only shows every fourth index, corresponding to one-year extensions of the sample). The top panel contains standard repeat-sales indices, and the bottom panel contains indices estimated with a simple trade-selection process containing just log LTV (Model A). In both panels, the consistently downward revisions are immediately apparent. For the selection-corrected indices, however, revisions are substantially smaller. For each quarter we calculate the standard deviation across the index values calculated for this quarter using the incrementally extended sample periods. For the standard repeat-sales index, the average within-quarter standard deviation is 0.060. For our selection-corrected index, it is 0.036.

Revisions are not only smaller, they also converge faster. To illustrate, Figure 7 plots the relative change in the index as the sample period is extended. The top plot contains revisions of the standard repeat-sales indices, and the bottom plot contains revisions of the selection-corrected ones. In both cases, the median revision when including a single ad-

[14] We considered two other external validity tests: (1) whether our model predicts prices of future out-of-sample transactions and (2) whether it predicts the timing of future out-of-sample trades and foreclosures. These tests are less useful, however. First, sample selection models do not attempt to predict observed outcomes (OLS is the best linear unbiased estimator), rather they estimate whether observed outcomes are representative of outcomes in the overall (unobserved) population. Second, our data contain no time-varying property-specific characteristics, making it difficult to predict the timing of individual trades and foreclosures.

ditional quarter of data is around -5%, although the magnitude of the revision is slightly smaller and the probability of a smaller revision is higher for the selection-corrected indices. In the limit, however, after 12 quarters of additional data are included, the median revision is around -13% for the standard repeat-sales index and around -8% for the selection-corrected index. In fact, this -8% median revision for the selection-corrected index is close to the "best case" revision (5th percentile) of -7% for the standard index. For the selection-corrected index, the best case revision is just -3%. Conversely, the "worst case" revision (95th percentile) for the standard index is as large as -21%. For the selection-corrected index, it is "only" -13%. Finally, Figure 7 illustrates the faster convergence of selection-corrected indices. These indices stabilize after 4 quarters of additional data, whereas the downward revisions of the standard repeat-sales index are present for the entire 12 quarters of additional data.

The reduced magnitude of the index-revision problem provides external validation of the model's ability to accurately capture the dynamic selection of traded properties. While the simple specification of the trade-selection process (Model A) does not fully eliminate the revisions, richer specifications of the selection processes may further reduce this problem. Due to the high computational cost of repeatedly estimating more complex specifications of the selection processes, we do not pursue this extension here.

IV LTV Distributions

The data contain the LTV ratio at the time of each trade. The top panel in Figure 2 plots the LTV ratio of the new owner's mortgage, and the bottom panel plots the seller's mortgage, constructed from amortizing the existing mortgages as described above. In Figure 8, the top panel plots the median seller's LTV over the sample period and the fraction of the housing stock sold in each quarter. The plots show an increase in median LTVs towards the end of the sample paired with a drop in trading volume. This increase in the median seller's LTV appears modest, essentially restoring the median LTV to its level before the housing boom during the mid-2000s. In contrast, the second Panel in Figure 8 reveals a substantial increase in the fraction of sales that are foreclosure sales. The bottom Panel

in Figure 8 shows an even greater increase in the fraction of properties sold with an LTV ratio above one, i.e., sales of underwater properties. This suggests that not only is the median LTV deteriorating, but the cross-sectional distribution of LTVs is also becoming more dispersed, with an increasing fraction of properties with very negative equity. These plots are made from the observed LTVs at the time of the transactions.

Turning to LTVs estimated by the model, Figures 9 to 11 plot historical LTV distributions in the population. Figure 9 shows the percentiles of the LTV distribution resulting from the various specifications, and Figures 10 and 11 show the fractions of properties with LTVs greater than 1, 1.25, and 1.5. In all cases, the LTV distribution appears to deteriorate substantially during the last years of our sample.

We compare the LTV distributions estimated using the model to those that would be obtained by imputing prices from a local house price index. In Figures 10 and 11 the plots denoted RSR represent LTV estimates constructed by imputing property prices using the standard repeat-sales index. For each property, the loan and price are recorded at the time when the property transacts, and updated prices are then calculated and imputed using the GLS index in Figure 3. The fraction of underwater properties according to this RSR calculation is substantially below both the fraction plotted for the baseline MCMC estimates—calculated without selection correction, but including the dispersion in prices arising from the idiosyncratic volatility of individual prices—and the two plots correcting for selection (using Models A and F). Compared to the RSR estimates, the baseline MCMC estimates show a substantial increase in the dispersion in LTVs, and consequently a greater fraction of underwater properties.

Figures 10 and 11 also show that moving from the baseline MCMC model to Models A and F leads to substantial changes in the estimated LTV distribution, which may be more surprising. This is due to a more subtle dynamic selection effect. When properties trade or are refinanced, their LTV is typically set to 80%. Appreciating properties have better (lower) LTVs but are also more likely to be refinanced and have their LTVs reset to 80%. Depreciating properties have worse (higher) LTVs but those are less likely to be reset. Hence, even when prices are constant, the average LTV may deteriorate due to this dynamic

selection effect.[15] This effect explains the increase in the fraction of underwater properties moving from the baseline model to Model A. Model F further allows the LTV coefficient to vary by period, and the coefficient changes sign in the last period, reversing the direction of the dynamic selection and resulting in a lower fraction of underwater properties than the baseline model. This selection of trades and refinancings, as a function on LTVs, is thus important for estimating the cross-sectional dispersion of prices and LTVs, particularly the fraction of underwater properties. In all cases, however, the RSR estimates using imputed prices underestimate this fraction.

V Trade and Foreclosure Behavior

Decisions to trade, prepay, and default are particularly important when pricing mortgage-backed securities. The literature models these decisions in two ways. First, decisions to trade and prepay can be viewed as exercising real options, and behavior can be derived from models of optimal exercise of options (e.g., Stanton (1995) and Longstaff (2005)). This structural approach, however, has been less successful empirically, either because homeowners' decisions are not optimal or because the models are too stylized to capture the nuances of these decisions. Alternatively, trade and prepayment intensities can be estimated from the observed hazard rates in pools of mortgages (e.g., Schwartz and Torous (1989, 1992)). One limitation of this reduced-form approach is that the trade and prepayment decisions are modeled as functions of only a limited set of explanatory variables. Importantly, these models do not permit these decisions to depend on contemporaneous price or LTV. As a consequence, hazard rates obtained this way tend to shift across vintages, suggesting that the parameters are not constant, structural parameters of the model, raising concerns about using these hazard rates to forecast future behavior and to price mortgage-backed securities.

Our model provides a new approach to estimating trade and foreclosure intensities. This is a reduced form approach where decisions to sell are captured by the trade-selection process, but this process is estimated at the individual property level (not from pools), allowing

[15] Jensen's inequality may produce an additional effect when calculating average LTVs by imputing average prices, because the LTV ratio is a convex function of the price, and $E\left[L/P|\delta\right] \geq L/E\left[P|\delta\right]$.

an arbitrary number of property-specific explanatory variables, including contemporaneous price or LTV.

A Trade Behavior

Table II presents the different specifications of the trade process. All models include log LTV as an explanatory variable, and Models B to E add additional variables. Model F allows the log-LTV coefficient to vary over the sample period to investigate changes in trade behavior.

In Model A in Table II, the estimated log-LTV coefficient is negative and statistically significant.[16] Since a lower LTV corresponds to a higher price, the negative coefficient suggests that properties with higher prices trade at higher intensities, consistent with Case, Pollakowski, and Wachter (1997), who also find that appreciating properties trade faster. Moreover, Figure 4 shows that the selection correction attenuates the size of the price bubble relative to the baseline model. With a negative log-LTV coefficient, this attenuation is intuitive: When prices generally appreciate, properties with more rapid price appreciation are traded more frequently, and their prices are observed more frequently, causing standard indices to exaggerate the price appreciation.

Model B in Table II, includes the time in years since the previous sale (*Time*). The coefficients for *Time* are positive and those for *Time-Squared* are negative, showing that the trade intensity follows an inverse U shape as a function of *Time*. When a property has just traded, the probability of another trade immediately drops and then gradually increases, peaking after about nine years, holding LTV constant.[17] As discussed, including *Time* also improves the statistical identification of the model, although Table II and Figure 4 suggest that this has a negligible effect on the estimated parameters and indices.

In Model C in Table II there is a negative and significant coefficient on the change in mortgage rate since loan inception, indicating that properties for which the mortgage rate has increased are less likely to trade, which is not surprising. Including this variable has little effect on the LTV coefficient, however, and Figure 4 shows almost no change in the

[16] Although a slight abuse of standard terminology for Bayesian statistics, we call a coefficient statistically significant at a given level when the corresponding credible interval does not contain zero.

[17] In Model B, the maximum intensity occurs after 9.1 years ($9.1 = -0.0381/(2 \times -0.0021)$).

index.

Model D in Table II includes the size and age of the property, and shows that larger and older houses trade less frequently. In Model E, these house characteristics are also interacted with the log-LTV ratio. The positive significant coefficient on the interaction of size and log LTV shows that for larger properties the trades are less sensitive to LTVs than they are for smaller houses. The interaction of age and log LTV shows a very small effect.

Finally, in Model F, the sample is divided into four sub-periods with separate intercepts and log-LTV coefficients. This specification is motivated by the sharp increase in foreclosures toward the end of the sample, which raises concerns about a structural break in trade and foreclosure behavior. Interestingly, the coefficient on LTV increases monotonically over the sample period (its absolute magnitude declines). The other coefficients are largely unchanged. The gradual increase in the LTV coefficients reflects a gradual shift in the market, where sales are becoming less sensitive to LTVs. In fact, over the last period the log-LTV coefficient changes sign, suggesting that recently, properties with higher LTVs have traded more frequently. One explanation for this change is the increase in foreclosure sales of underwater properties. Indeed, Figure 4 shows that the price index estimated with the specification of the trade process given by Model F declines by about 10–15 index points less during 2007 and 2008 than the other indices. Moreover, since the index from Model F peaks at a higher level, the proportional decline is even smaller. We investigate this shift in more detail below, when we estimate separate processes for normal trades and foreclosure sales. Another explanation for the change in trading behavior would be a change in economic conditions, resulting in a change in the demand for inexpensive, high-LTV housing. We re-estimate (but do not report) our model with data for Maricopa County, Arizona (Phoenix and Scottsdale metro areas) and Clark County, Nevada (Greater Las Vegas area), and find qualitatively similar results. Specifically the dramatic deterioration of the LTV distributions and the gradual increase in the LTV coefficients over the sample period appear to be robust across these geographical locations. If the changes in the trading behavior are due to economic conditions, these economic conditions appear to be widely shared.

So far, the estimates only include properties that trade (at least) twice to make the

analysis comparable to the standard repeat-sales analysis. One may be concerned this sub-sample is selected, although this is probably a smaller concern given the long sample period. Unlike the traditional repeat-sales regression, our model can be estimated using all properties in the data, including those that trade only once, assuming they follow the same trade process as those properties that trade more than once. Including these properties roughly doubles the sample size to over 140,000 properties, which may improve the estimates of the price and trade processes (mechanically, the level of the estimated trade intensity is reduced). Table III compares the coefficient on LTV in the trade equation across the samples. In the extended sample, the coefficient is closer to zero, but still significant. Another concern is that foreclosure sales may not be arms-length sales and that the resulting prices may not reflect the true market values. We re-estimate our model after eliminating foreclosure sales from the sample to investigate this concern, but Table III shows that the LTV coefficient remains largely unchanged.

B Foreclosure Behavior

Table IV presents coefficients for specifications with separate processes for foreclosure sales and normal trades. Panel A contains the trade process, and Panel B contains the foreclosure process. The coefficients in Panel A are largely similar to those from Table II, which combines normal trades and foreclosure sales in the estimation. This follows from the relatively low number of foreclosures over most of the sample. Hence, normal trades dominate the sample and estimates using just normal trades appear very similar to estimates that combine the normal trades with the (relatively few) foreclosures sales.

Comparing Panels A and B in Table IV, the greatest difference between the trade and foreclosure processes is the LTV coefficient for the last sub-period in Model F. In Model F in Panel A, the coefficient on LTV in the last sub-period is smaller compared to Table II. Much of this increase appears to be caused by foreclosures of high-LTV properties. This is confirmed in Panel B in Table IV, where Model F has a large positive coefficient on LTVs in the last period, showing that these foreclosure sales were predominantly driven by properties with high LTVs. The negative coefficients during the previous periods are less intuitive, but these are periods with fewer foreclosure sales, so the estimate may be a result

of more atypical transactions.

Comparing Panels A and B in Table IV more broadly, the trade and foreclosure processes appear largely similar. In addition to the differences for log LTV, the two processes appear to respond slightly differently to property age, with older properties more likely to foreclose and less likely to trade in normal trades. Nevertheless, Figures 4 and 5 show that the indices are largely unchanged after including these processes separately, despite their differences.

C Equivalent Constant Intensity

To interpret the economic magnitudes of the coefficients in the trade equation, we define the *equivalent constant intensities* as follows. Assuming that trades follow a Poisson arrival process with constant intensity λ_i, the number of trades of property i up to time t, denoted $d_i(t)$, is distributed as:

$$P[d_i(t) - d_i(t-1) = n] = \frac{\lambda_i^n \exp[-\lambda_i]}{n!} \text{ for } n \geq 0. \tag{12}$$

The probability of no trade between time $t-1$ and t is:

$$P[d_i(t) - d_i(t-1) = 0] = \exp[-\lambda_i]. \tag{13}$$

Using the estimated coefficients in the trade process, the probability of no trade is:

$$P[w_i(t) < 0] = \Phi[-W_i'(t)\alpha_0 - p_i(t)\alpha_p], \tag{14}$$

where Φ is the *c.d.f.* of the standard normal distribution. Equating these probabilities,

$$\exp[-\lambda_i] = \Phi[-W_i'(t)\alpha_0 - p_i(t)\alpha_p], \tag{15}$$

implying:

$$\lambda_i = -\ln\left[\Phi[-W_i'(t)\alpha_0 - p_i(t)\alpha_p]\right], \tag{16}$$

with partial derivative:

$$\frac{d\lambda_i}{dW_i} = \alpha_0 \frac{\phi[-W_i'(t)\alpha_0 - p_i(t)\alpha_p]}{\Phi[-W_i'(t)\alpha_0 - p_i(t)\alpha_p]}. \tag{17}$$

We call λ_i the equivalent constant intensity. This intensity is the constant Poisson intensity that gives rise to the same probability of a trade as the estimated coefficients in the trade process, and its partial derivatives are more readily interpretable as economic magnitudes than the estimated coefficients. This is closely analogous to marginal effects for probit models; in fact, the mathematical expressions are quite similar. The posterior distributions of λ_i and the partial derivatives are simple to calculate from the posterior distribution of α. Note that λ_i is only used to interpret the economic magnitude of the coefficients in the trade process. It is not a structural part of the model. Explicitly modeling the trade intensity as a continuous-time process depending on the LTV (or price) turns it into a *doubly stochastic Poisson process* or Cox process (see Cox, 1955), which is beyond the present scope.

Figure 12 plots the time series of λ_i for Models A and F, evaluated at the median LTV and the 25th and 75th percentiles across properties. The top panels are from the specifications with only a trade process, and the bottom panels are from the specifications with separate trade and foreclosure processes. The top plots show slightly higher intensities, because they plot the combined intensities of either normal trades or foreclosure sales. The bottom panels are just the intensities of normal trades. Moving from Model A in the left panels to Model F on the right, the intensity process becomes much more volatile. Model F allows the LTV coefficient to vary over the sample period, and it includes a greater number of explanatory variables, including mortgage rates and quarterly indicators to capture seasonality in the trade process, overall capturing more time variation in the process. Specifically, the seasonal adjustments produces the "spiky" movements in the intensity.

To interpret the economic magnitude of the estimated LTV coefficients, the thin gray lines in Figure 12 plot the intensities evaluated at the bottom and top quartiles of the LTV distribution. In Model A, the LTV coefficient combined with the widening of the LTV distribution shows that early in the sample period, the trade intensities for relatively high- and low-LTV properties were quite similar, around 2.4%. Later in the sample period, the intensities diverge, with a quarterly trade intensity of low-LTV properties around 2.25% and high-LTV around 2.75%, representing about a 20% difference. In Model F, the additional

explanatory variables lead to a less clear picture. Not surprisingly, we find that the trade intensity for normal trades increases in the early 2000s, and decline rapidly towards the end of the sample period.

Turning to foreclosure intensities, Figure 13 reveals the substantial increase in these intensities in recent years, along with a quiet period during the price run-up in the early 2000s preceded by substantial activity during the previous housing crisis in the mid 1990s.

Table V reports the partial derivatives of the equivalent constant intensity. Panel A shows the effect on the trade intensity. From Figure 12 we know that the average quarterly intensity is around 2.5%, and the mean estimate of -0.49% shows that a one-standard deviation increase in LTV reduces the trade intensity from around 2.5% to around 2% (around a 20% decline), which confirms that the economic magnitude of the effect of LTV on trade intensity is substantial. The corresponding figures for foreclosures are in Panel B. The first figure of 0.04% shows that a one-standard deviation increase in LTV is associated with an increase in foreclosure intensity from around 0.1% (from Figure 13) to 0.14%, a substantial increase. Finally, note that these figures mechanically overestimate the baseline intensities, because only properties that trade twice (or more) are included in the sample.

VI Conclusion

While price indices, LTV distributions, and trade and foreclosure behavior are common indicators of economic activity, they are surprisingly difficult to estimate, because property prices are only observed when the properties trade, and trades are endogenous. We present a new econometric model to address this problem. Our approach filters out the entire price path of each property in our data, even when the property is non-traded and its price is unobserved. This filtering is numerically intensive, but exploiting recent advances in computational Bayesian estimation —specifically, MCMC, Gibbs sampling, and FFBS— produces a tractable estimation procedure with robust convergence properties.

The model estimates property prices jointly with the trade and foreclosure processes, which allows those decisions to depend directly on price or LTV. It uses the trade and foreclosure processes as dynamic extensions of the standard cross-sectional Heckman sample-

selection equation to correct the estimates for selection arising when the price dynamics for the traded properties (for which prices are observed) are not representative of the overall population.

For price indices, there have been concerns that the recent increase in the fraction of foreclosure sales biases standard repeat sales indices. We find that the magnitude of this bias is more modest than expected, at least in the long term. Even when rapidly appreciating properties were more likely to trade during the bubble years, and this created an upward bias in the traditional repeat-sales index, this bias appears to be somewhat modest in magnitude and temporary as well. When the more slowly appreciating properties eventually trade, they bring the index back in line with the average price level in the underlying population.

In the short term, however, selection leads to the index revision problem where current index estimates are subsequently revised (typically downwards) as trade data become available for the estimation. These revisions are substantial, but our baseline specification reduces the magnitudes of the revisions by about 45% (the absolute magnitude of the median revision declines from 13% to 7%).

We find large effects for home price dispersions, compared to an estimator of LTV distributions that imputes prices from a repeat-sales index on the untraded properties for which the price is unobserved. This standard approach substantially underestimates the dispersion of prices, and hence the dispersion of LTVs, leading to estimates of the fraction of underwater properties substantially below those produced by our approach.

Finally, we calculate the equivalent continuous-time Poisson intensities. These are potentially useful for models pricing mortgage-backed securities and derivatives on home price indices. Moreover, these intensity estimates suggest large recent changes in trade and foreclosure behavior. While we cannot isolate the specific cause of this change, it appears to be wide spread, and some caution is required when using historical data to predict future prepayment and defaults

To our knowledge, this study contains the first formal econometric analysis of the problem of estimating LTV distributions, given the illiquidity of the real estate market. More research is needed, however. Our sample ends in 2008, and while we have been unable

to extend the sample beyond this period, the deterioration of the real estate markets has continued. In April, 2010, Fiserv (the publisher of the S&P/Case-Shiller Indices) stated that "the housing market has experienced significant turmoil and the last two-to-three years have seen large increases in foreclosures as well as other market dislocations" and followed this statement by a recommendation against using the standard seasonality adjustments, as they appear to no longer work correctly. More recently, the June 2011 home sales figures from Las Vegas show that volume is up 8% year-over-year but that almost 70% of sales were distressed sales (47.2% were bank-owned properties and 21.6% were short sales), underscoring the importance of understanding distressed markets for illiquid assets.

Appendix: Estimation Procedure

For each property i, the natural logarithm of price, $p_i(t)$, follows the process:

$$p_i(t) = p_i(t-1) + \delta(t) + \varepsilon_i(t), \tag{18}$$

with $\varepsilon_i(t) \sim N(0, \sigma^2)$ and *i.i.d.* across firms, $i = 1 \ldots N$, and across time, $t = 1 \ldots T$. Note that $\delta(t)$ is the index return between time $t-1$ and t, so $\delta(t)$ ranges from $\delta(2)$ to $\delta(T)$. The price is observed whenever $w_i(t) \geq 0$, where $w_i(t)$ is given by the selection equation:

$$w_i(t) = W_i'(t)\alpha_0 + p_i(t)\alpha_p + \eta_i(t). \tag{19}$$

The vector of covariates $W_i(t)$ is observed. In some models, we separate foreclosure sales from normal trades by including an additional selection equation in which a foreclosure occurs when $z_i(t) \geq 0$, where $z(t)$ is:

$$z_i(t) = Z_i'(t)\gamma_0 + p_i(t)\gamma_p + \xi_i(t). \tag{20}$$

The error terms η and ξ are distributed *i.i.d.* normal with variance equal to one, and are uncorrelated with each other at all leads and lags (including contemporaneously).

It is convenient to stack the selection equations into a 2x1 vector:

$$s_i(t) = S_i'(t)\phi_0 + p_i(t)\phi_p + \zeta_i(t), \tag{21}$$

where $s_i(t) = \begin{bmatrix} w_i(t) \\ z_i(t) \end{bmatrix}$, $S_i'(t) = \begin{bmatrix} W_i'(t) & 0 \\ 0 & Z_i'(t) \end{bmatrix}$, $\phi_0 = \begin{bmatrix} \alpha_0 \\ \gamma_0 \end{bmatrix}$, $\phi_p = \begin{bmatrix} \alpha_p \\ \gamma_p \end{bmatrix}$, and $\zeta_i(t) = \begin{bmatrix} \eta_i(t) \\ \xi_i(t) \end{bmatrix}$. The covariance matrix of $\zeta_i(t)$ is the identity matrix. For the models in which we only use one selection equation, $s_i(t) = w_i(t)$ and equation (21) is identical to (19).

The set of parameters to be estimated is $\theta = \left(\delta(2) \ldots \delta(T), \sigma^2, \alpha_0, \alpha_p, \gamma_0, \gamma_p\right)$. We augment the parameter set with the latent variables and use a Bayesian estimation algorithm

that simulates the posterior distribution, $f(\theta, \{p_i(t), s_i(t)\} | data)$, using a Gibbs sampler (Gelfand and Smith, 1990). By the Hammersley-Clifford theorem, we can break up the posterior into three *complete conditionals*:

1. Latent prices: $f(\{p_i(t)\} | \{s_i(t)\}, \theta, data)$
2. Selection variables: $f(\{s_i(t)\} | \{p_i(t)\}, \theta, data)$
3. Parameters: $f(\theta | \{p_i(t), s_i(t)\}, data)$

We sample from each distribution 1-3 in turn, after which we return back to step 1 and repeat. The resulting sequence of parameter draws forms a Markov chain, the stationary distribution of which is exactly the posterior distribution. Given a sample of draws of the posterior distribution, it is then straightforward to numerically integrate out the latent variables and obtain the marginal posterior of parameters, $f(\theta|data)$, or the unobserved prices, $f(\{p_i(t)\} | data)$, for example. We now discuss how to draw from each conditional distribution.

A1 Latent prices

We draw latent prices in the period between property sales using the FFBS algorithm (Carter and Kohn, 1994; and Fruhwirth-Schnatter, 1994), which provides an efficient way to sample a path of state variables defined by a linear state space model. Since the error terms are assumed *i.i.d.* across properties, we can sample $p_i(t)$ separately for each propety. For expositional simplicity, we describe the algorithm for a particular property, suppressing the dependence on i.

Interpreting the econometric model as a linear state space model, $p(t)$ is the state variable, and equation (18) is the transition rule. Conditional on the parameters, $\delta(t)$ is an "observed" control acting on the state, and conditional on $s(t)$, the collection of selection equations given by equation (21) are noisy observations equation for the state. This setup allows us to calculate the filtered distribution of $p(1) \dots p(T)$, using the Kalman filter.

The Kalman filter produces the distribution of $p(t)$ conditional on $s(1) \dots s(t)$, for any time t. However, $p(t)$ needs to be sampled conditional on the entire time series $s(1) \dots s(T)$.

This is achieved by a backward smoother, which effectively runs a Kalman filter backwards, starting at time T. The conditional distribution of the state vector of latent valuations is given by the following identity, which follows from Lemma 2.1 in Carter and Kohn (1994):

$$f\left(p(1)\dots p(T)|s^T\right) = f\left(p(T)|s^T\right)\prod_{t=1}^{T-1} f\left(p(t)|s^t, p(t+1)\right), \tag{22}$$

where $s^t = \{s(1)\dots s(t)\}$ contains the selection variables up to time t. Next we describe the forward filtering and backward sampling steps in detail.

Define $m(t|j) = E\left[p(t)|s^j\right]$ and $v(t|j) = Var\left[p(t)|s^j\right]$ as the mean and variance of $p(t)$ conditional on the selection variables up to time j. Note that all conditional distributions are normal and hence fully characterized by their means and variances (Kalman, 1960; and Anderson and Moore, 1979).

For the forward filtering step, for $t = 1\dots T$, we calculate $m(t|t)$ and $v(t|t)$ by iterating on the forward filter, through a forecasting and an updating part. The forecasting part involves the two equations:

$$m(t|t-1) = m(t-1|t-1) + \delta(t), \tag{23}$$

and

$$v(t|t-1) = v(t-1|t-1) + \sigma^2. \tag{24}$$

For the updating part, as long as $p(t)$ remains unobserved, we update:

$$m(t|t) = m(t|t-1) + K' \cdot \left[s(t) - S'(t)\phi_0 - m(t|t-1)\phi_p\right], \tag{25}$$

where the Kalman gain K is given by:

$$K = \left[I + \phi_p v(t|t-1)\phi_p'\right]^{-1} \cdot v(t|t-1)\phi_p, \tag{26}$$

and I is the 2x2 identity matrix (in the case of one selection equation, this is simply a scalar unity). When K is large, more weight is placed on the information from the selection equation. This happens when either ϕ_p or $v(t|t-1)$ is large, i.e., when either the selec-

tion equations are more informative about the valuations or when the valuations are more uncertain. Further,

$$v(t|t) = v(t|t-1) \cdot \left(1 - K' \cdot \phi_p\right). \tag{27}$$

We force $\phi_p = 0$ when estimating the model without correcting for selection. Then, $m(t|t) = m(t|t-1)$ and $v(t|t) = v(t|t-1)$, and no information is used in periods where $p(t)$ is unobserved. In periods where $p(t)$ is observed, $m(t|t) = p_t^{OBS}$ and $v(t|t) = 0$.

For the backward sampling part, $p(T)$ is first simulated from the normal distribution with mean $m(T|T)$ and variance $v(T|T)$, as given by the Kalman filter. For $t = T-1 \ldots 1$, we draw $p(t)$ from the conditional distribution $p(t)|s^t, p(t+1)$. This distribution can be derived from a filtering problem where the draw of $p(t+1)$ provides an additional observation of $p(t)$. The distribution is:

$$p(t)|s^t, p(t+1) \sim \mathcal{N}(r,q), \tag{28}$$

where:

$$r = m(t|t) + G \cdot [p(t+1) - m(t+1|t)], \tag{29}$$

$$q = v(t|t) \cdot (1-G), \tag{30}$$

with:

$$G = \frac{v(t|t)}{v(t|t) + \sigma^2}. \tag{31}$$

From equation (31), G can be interpreted as a Kalman gain similar to K in equation (26). As such, the backwards sampler weighs the information from the filtered distribution $p(t)|s^t$ and the information in $p(t+1)|s^T$ to obtain a draw of $p(t)|s^T$, with the weight depending on the relative variance of the filtered estimate, $v(t|t)$, and the variance of a one-period price change. If the filtered estimate $m(t|t)$ is very precise relative to the variance of the valuation change from one period to the next, then G is close to zero, and most of the weight is put on the distribution of $p(t)$ from the Kalman filter. The more imprecise the Kalman filter distribution relative to how much the valuation can possibly change (as captured by sigma), the more weight is put on the "observed" $p(t+1)$.

A2 Selection variables

The selection variables are sampled conditional on the valuations, parameters, and whether or not the valuation is observed. Simulating this block is similar to simulating the (augmented) posterior distribution of a probit model (Albert and Chib, 1993). Under the assumption that η and ξ are independent, we may draw the selection variables, w and z, separately. In addition, by the *i.i.d.* assumption we may draw each property-quarter variable separately.

When property i is sold, the price is observed and the posterior distribution of the first selection variable, $w_i(t)$, is:

$$w_i(t) | \{p_i(t)\}, \theta, data \sim \mathcal{N}_L\left(W_i'(t)\alpha_0 + p_i(t)\alpha_p, 1\right). \tag{32}$$

When home price is unobserved, the distribution is:

$$w_i(t) | \{p_i(t)\}, \theta, data \sim \mathcal{N}_U\left(W_i'(t)\alpha_0 + p_i(t)\alpha_p, 1\right). \tag{33}$$

Here, $\mathcal{N}_L\left(\mu, \sigma^2\right)$ denotes a normal distribution with mean μ and variance σ^2 that is truncated below at zero. Similarly, $\mathcal{N}_U\left(\mu, \sigma^2\right)$ is the upper-truncated distribution, truncated above at zero.

Drawing $z_i(t)$ is analogous but using foreclosures instead of regular home sales for observations.

A3 Parameters

Conditional on $\{p_i(t)\}$, $\{W_i(t)\}$, and $\{Z_i(t)\}$, the distributions of α, γ, $\{\delta(t)\}$, and σ^2 are given by the three Bayesian linear regressions (18), (19), and (20). Since ε, η, and ξ are independent by assumption, we may estimate the three equations separately.

In the valuation equation, $\delta = [\delta(2) \ldots \delta(T)]'$ and σ^2 are defined by the regression of Y_p on X_p, where the vector Y_p stacks the one-period returns, $p_i(t) - p_i(t-1)$, across all properties and time periods. Let $N(t)$ be the number of companies for which $p(t) - p(t-1)$ exists, so Y_p is a $\sum_{t=2}^{T} N(t)$ by 1 vector. The matrix X_p is a $\sum_{t=2}^{T} N(t)$ by $T-1$ matrix of

zeros and ones. Each row of X_p contains $T-2$ zeros and a one in column $t-1$, corresponding to the timing of the return in Y_p (such that a one in the first column of X_p indicates a return from time 1 to time 2).

The standard conjugate normal inverse gamma prior with prior parameters a_0,b_0, μ_0, and Σ_0 is:

$$\sigma^2 \sim IG(a_0, b_0), \tag{34}$$

$$\delta|\sigma^2 \sim \mathcal{N}\left(\mu_0, \sigma^2\Sigma_0^{-1}\right). \tag{35}$$

The posterior distributions for the parameters in the valuation equation are then (e.g., Rossi, Allenby, and McCulloch, 2005):

$$\sigma^2|Y_p, X_p \sim IG(a, b), \tag{36}$$

$$\delta|\sigma^2, Y_p, X_p \sim \mathcal{N}\left(\mu, \sigma^2\right), \tag{37}$$

with parameters:

$$a = a_0 + \sum_{t=2}^{T} N(t), \tag{38}$$

$$b = b_0 + e'e + (\mu-\mu_0)'\Sigma_0(\mu-\mu_0), \tag{39}$$

$$\Sigma = \Sigma_0 + X_p'X_p, \tag{40}$$

$$\mu = \Sigma^{-1}\left(\Sigma_0\mu_0 + X_p'Y_p\right). \tag{41}$$

The vector $e = Y_p - X_p\mu$ contains the stacked error terms.

The above regression pools the entire panel data set and quickly leads to memory issues and slow computation speeds. For example, in a data set with 100,000 properties observed for 80 quarters, the X_p matrix is 8 million by 79. The solution to these problems is to exploit the unique structure of X_p. In particular, $X_p'X_p$ is a diagonal $T-1$ by $T-1$ matrix that contains the number of trades on the diagonal. With a diagonal prior Σ_0, the inverse Σ^{-1} is also a diagonal matrix with the inverse of each element of Σ on the diagonal. The $T-1$ by 1 vector $X_p'Y_p$ contains the sum of returns for each period. These quantities can

be efficiently computed and used in equations (38)-(41).

The selection equations are considerably simpler. To obtain a draw of $\alpha = \begin{bmatrix} \alpha_0 \\ \alpha_p \end{bmatrix}$, we regress Y_w on X_w, where Y_w is a vector that stacks $w_i(t)$ over all properties and time periods. Similarly, X_w stacks $\begin{bmatrix} W_i'(t) & p_i(t) \end{bmatrix}$ over all properties and periods. Recall that we normalize the variance of the error term to one in order to identify the scale of the parameters, and we can consequently treat the inference problem as a standard Bayesian regression with known variance. The prior distribution is:

$$\alpha \sim \mathcal{N}\left(\theta_0, \Omega_0^{-1}\right), \tag{42}$$

and the posterior becomes:

$$\alpha | Y_w, X_w \sim \mathcal{N}\left(\theta, \Omega^{-1}\right), \tag{43}$$

with:

$$\Omega = \Omega_0 + X_w' X_w, \tag{44}$$

and:

$$\theta = \Omega^{-1}\left(\Omega_0 \theta_0 + X_w' Y_w\right). \tag{45}$$

Drawing γ works analogously to drawing α.

A4 Priors and Starting Values

Our Gibbs sampler uses 1,000 iterations for the initial burn-in, followed by 500 iterations to simulate the posterior distribution. During the burn-in, the simulations converge quickly. We use diffuse priors for the parameters. The prior distribution of σ^2 is inverse gamma with parameters $a_0 = 2.1$ and $b_0 = 1/100$, implying that $E[\sigma] = 8.5\%$ per quarter, and σ is between 3.6% and 28.6% (quarterly) with 99% probability.

We set prior means for α and δ to zero ($\theta_0 = 0$ and $\mu = 0$). We set $\Omega_0^{-1} = I/100$, where I is the identity matrix, so that the α are between -20 and +20 with 95% probability. We assume that $\Sigma_0^{-1} = I/100$. Together with the prior on σ^2, this implies that our prior on the one-quarter log change in price index is between -170% and +170% with 95% probability.

We start the algorithm with α and δ equal to zero, and $\sigma = 25\%$. We do not need starting values for the missing house prices or the selection variables, since the missing house prices are the first variable we simulate, and do not depend on *w* because we start with $\alpha = 0$.

When we use two selection equations (separating regular transactions from and foreclosure sales), we assume the same prior distribution and starting values for the coefficients of the two selection equations.

We implement this algorithm in C++, using the GNU Scientific Library (GSL). On a desktop 2.66 GHz Pentium 4 quad-core processor, it takes anywhere from less than one hour for the simpler models to about five hours for the most complex model to simulate 1,500 draws of the Markov Chain (using only a single core).

References

Albert, James and Siddhartha Chib, 1993, Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association* 88, 669–679.

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse, 1963, A Regression Method for Real Estate Price Index Construction, *Journal of the American Statistical Association* 58, 933–942.

Bajari, Patrick, Sean Chu, and Minjung Park, 2010, An Empirical Model of Subprime Mortgage Default from 2000 to 2007, *Working Paper*, University of Minnesota.

Bhutta, Neil, Shan, Hui and Dokko, Jane K., 2010, The Depth of Negative Equity and Mortgage Default Decisions, *Working Paper*, FEDS Working Paper No. 2010-35.

Calnea Analytics, 2010, Land Registry House Price Index Methodology: Discussion and Description of Methodology, http://www.calnea.net/LandRegistryIndex.pdf, downloaded March 3, 2010.

Carter, Chris K. and Robert J. Kohn, 1994, On Gibbs Sampling for State Space Models, *Biometrika* 81, 541–553.

Case, Karl E., Henry O. Pollakowski, and Susan M. Wachter, 1997, Frequency of Transaction and House Price Modeling, *Journal of Real Estate Finance and Economics* 14, 173–187.

Case, Karl E. and Robert J. Shiller, 1987, Prices of Single Family Homes since 1970: New Indexes for Four Cities, *New England Economic Review* Sept/Oct, 46–56.

Case, Karl E. and Robert J. Shiller, 1989, The Efficiency of the Market for Single-Family Homes, *American Economic Review* 79, 125–137.

Clapham, Eric, Peter Englund, John M. Quigley, and Christian L. Redfearn, 2006, Revisiting the Past and Settling the Score: Index Revisions for House Price Derivatives, *Real Estate Economics* 34, 275–302.

Cochrane, John, 2005, The risk and return of venture capital , *Journal of Financial Economics* 75, 3–52.

Cox, D. R., 1955, Some statistical methods connected with series of events, *Journal of the Royal Statistical Society, Series B* 17, 129–164.

Downing, Chris, Richard Stanton and Nancy Wallace, 2005, An Empirical Test of a Two-

Factor Mortgage Valuation Model: How Much do House Prices Matter?, *Real Estate Economics* 33, 681–710.

Ferreira, Fernando, Joseph Gyourko and Joseph Tracy, 2010, Housing Buses and Household Mobility, *Journal of Urban Economics* 68, 34–45.

Fruhwirth-Schnatter, Sylvia, 1994, Data Augmentation and Dynamic Linear Models, *Journal of Time Series Analysis* 15, 183–202.

Gatzlaff, Dean H. and Donald R. Haurin, 1997, Sample Selection Bias and Repeat-Sales Index Estimates, *Journal of Real Estate Finance and Economics* 14, 33–50.

Gatzlaff, Dean H. and Donald R. Haurin, 1998, Sample Selection and Biases in Local House Value Indices, *Journal of Urban Economics* 43, 199–222.

Gelfand, Alan and Adrian Smith, 1990, Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association* 85, 398–409.

Geman, Stuart and Donald Geman, 1984, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.

Genovese, David and Christopher Mayer, 2001, Loss Aversion and Seller Behavior: Evidence from the Housing Market, *Quarterly Journal of Economics* 116, 1233–1260.

Goetzmann, William N., 1992, The Accuracy of Real Estate Indices: Repeat Sale Estimators, *Journal of Real Estate Finance and Economics* 5, 5–53.

Goetzmann, William N. and Liang Peng, 2006, Estimating House Price Indexes in the Presence of Seller Reservation Prices, *Review of Economics and Statistics* 88, 100–112.

Goetzmann, William N. and Matthew Spiegel, 1995, Non-Temporal Components of Residential Real Estate Appreciation, *The Review of Economics and Statistics* 77, 199–206.

Haurin, Donald R. and Patric H. Hendershott, 1991, House Price Indexes: Issues and Results, *AREUEA Journal* 19, 259–269.

Heckman, James, 1979, Sample Selection Bias as a Specification Error, *Econometrica* 47, 153–162.

Heckman, James, 1990, Varieties of Selection Bias, *American Economic Review* 80, 313–318.

Himmelberg, Charles, Chris Mayer, Todd Sinai, 2005, Assessing High House Prices: Bubbles, Fundamentals, and Misperceptions, *Journal of Economic Perspectives* 19, 67–92.

Hwang, Min and John M. Quigley, 2004, Selectivity, Quality Adjustment and Mean Reversion in the Measurement of House Values, *Journal of Real Estate Finance and Economics* 28, 161–178.

Johannes, Michael and Nick Polson, 2006, MCMC Methods for Financial Econometrics, in Yacine Ait-Sahalia and Lars P. Hansen, eds.: Handbook of Financial Econometrics.

Jud, G. Donald, and Terry G. Seaks, 1994, Sample Selection Bias in Estimating Housing Sales Prices, *Journal of Real Estate Research* 9, 289–298.

Kain, John F. and John M. Quigley, 1970, Measuring the Value of Housing Quality, *Journal of the American Statistical Association* 65, 532–548.

Korteweg, Arthur G., and Morten Sørensen, 2010, Risk and Return Characteristics of Venture Capital-backed Entrepreneurial Companies, *Review of Financial Studies* 23, 3738-3772.

Korteweg, Arthur G., 2012, Markov Chain Monte Carlo methods in Corporate Finance, in P. Damien, P. Dellaportas, N. Polson, and D. Stephens, eds.: MCMC and Hierarchical Models.

Landvoigt, Tim, Monika Piazzesi, Martin Schneider, 2011, The Housing Market(s) of San Diego, *Working Paper*, Stanford University.

Landier, Augustin Landier, David Sraer and David Thesmar, 2011, The Risk-Shifting Hypothesis: Evidence from Subprime Mortgage Originations, *Working Paper*, Princeton University.

Laufer, Steven, 2011, Equity Extraction and Mortgage Default, *Working Paper*, NYU.

Lin, Zhenguo, and Kerry D. Vandell, 2007, Illiquidity and Pricing Biases in the Real Estate Market, *Real Estate Economics* 35, 291–330.

Longstaff, F. A., 2005, Borrower Credit and the Valuation of Mortgage-Backed Securities, *Real Estate Economics* 33, 619–661.

Meese, Richard A. and Nancy E. Wallace, 1997, The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression, and Hybrid Approaches, *Journal of Real Estate Finance and Economics* 14, 51–73.

Melzer, Brian, 2011, Mortgage Debt Overhang: Reduced Investment by Homeowners with Negative Equity, *Working Paper*, Northwestern University.

Munneke, Henry J. and Barrett A. Slade, 2000, An Empirical Study of Sample-Selection Bias in Indices of Commercial Real Estate, *Journal of Real Estate Finance and Economics* 21, 45–64.

Munneke, Henry J. and Barrett A. Slade, 2001, A Metropolitan Transaction-Based Commercial Price Index: A Time-Varying Parameter Approach, *Real Estate Economics* 29, 55–84.

Okah, Ebiere and Orr, James, 2010, Subprime Mortgage Lending in New York City: Prevalence and Performance, *Working Paper*, Federal Reserve Bank of New York Staff Report No. 432.

Rossi, Peter E., Greg M. Allenby and Robert McCullough, 2005, Bayesian Statistics and Marketing, John Wiley and Sons, Chichester, UK.

Schulhofer-Wohl, Sam, 2011, Negative Equity Does Not Reduce Homeowners' Mobility, *Working Paper*, Federal Reserve Bank of Minneapolis.

Schwartz, Eduardo and W. Torous, 1989, Prepayment and the Valuation of Mortgage-Backed Securities, *Journal of Finance* 44, 375–392.

Schwartz, Eduardo and W. Torous, 1992, Prepayment, Default, and the Valuation of Mortgage-Backed Securities, *Journal of Business* 65, 221–239.

Stanton, Richard, 1995, Rational Prepayment and the Valuation of Mortgage-Backed Securities, *Review of Financial Studies* 8, 677–708.

Tanner, Martin and Wing Wong, 1987, The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association* 82, 528–549.

Figure 1: Descriptive Statistics of Home Transactions. The figure shows the distribution of trades for Alameda County, California. Panel A presents the total number of trades, separating normal transactions and foreclosure sales. Panel B presents foreclosures as a fraction of all sales. Panels C and D present regular transactions and foreclosure sales as a fraction of the total number of properties in the data.

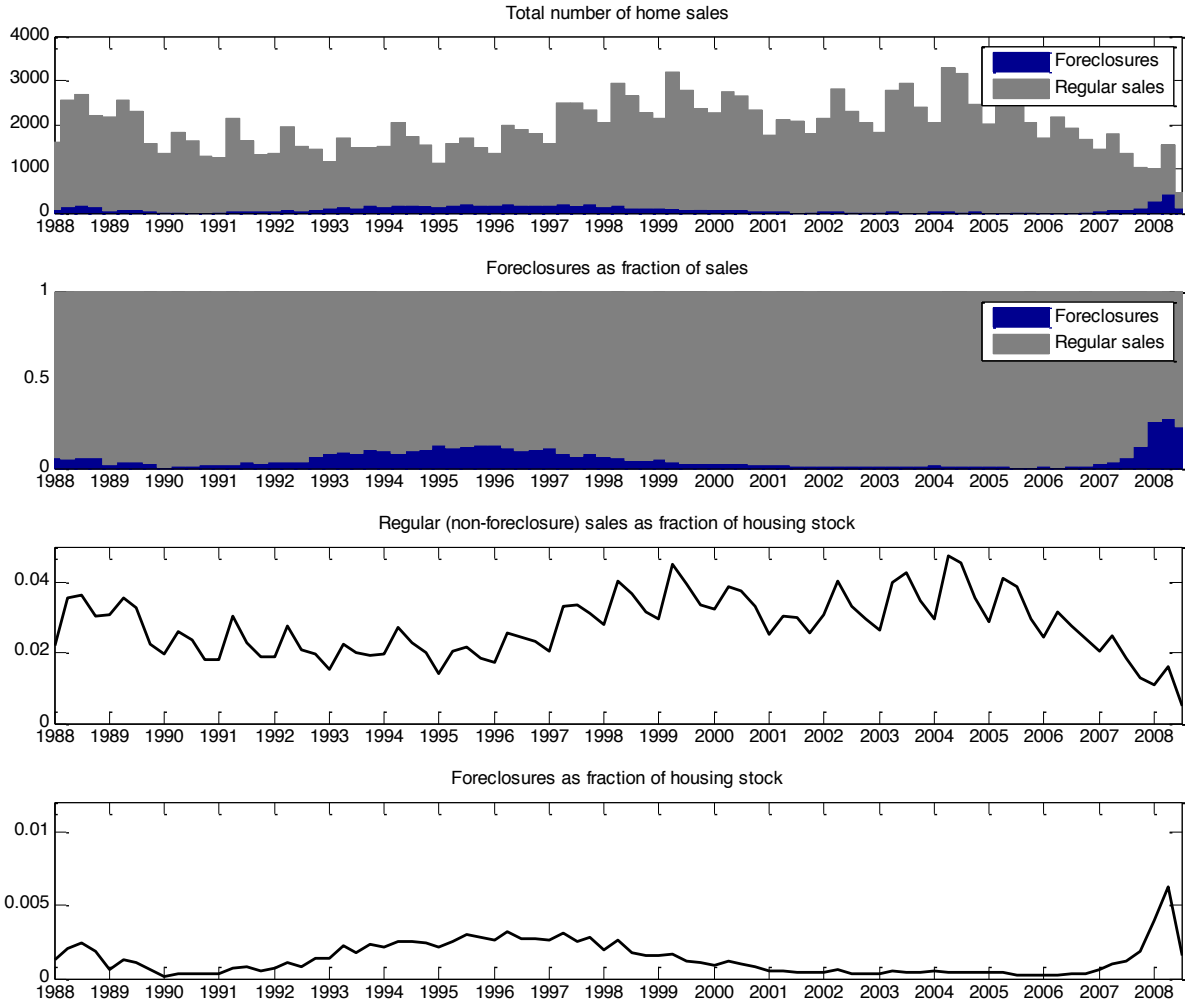
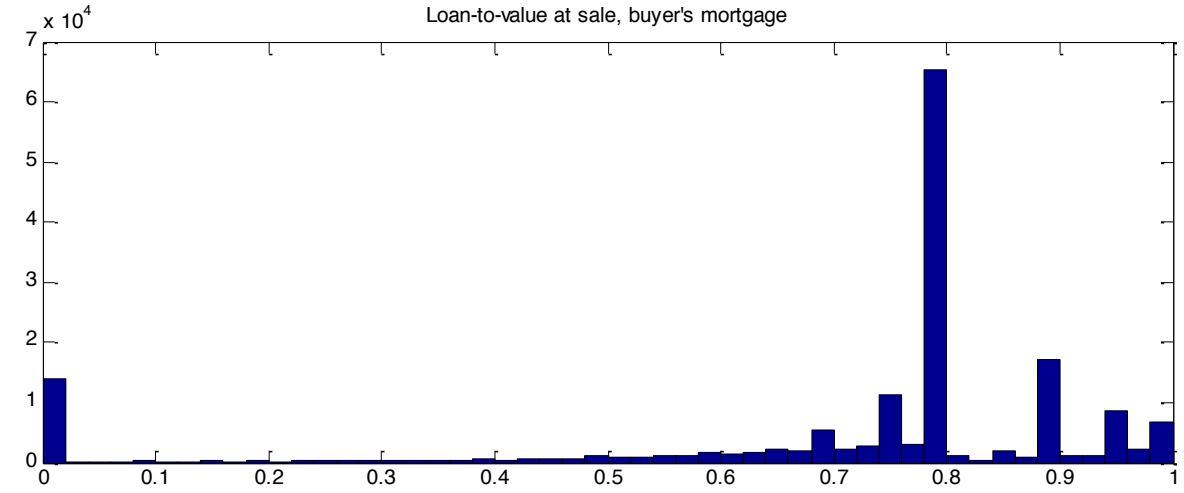


Figure 2: Distribution of LTV. The figure shows the distribution of buyers and sellers' LTVs at the time of sales. The top plot shows the histogram of LTVs where the loan amount represents the buyer's mortgage amount. The bottom plot uses the seller's remaining mortgage balance (computed as described in the text) to calculate LTVs.



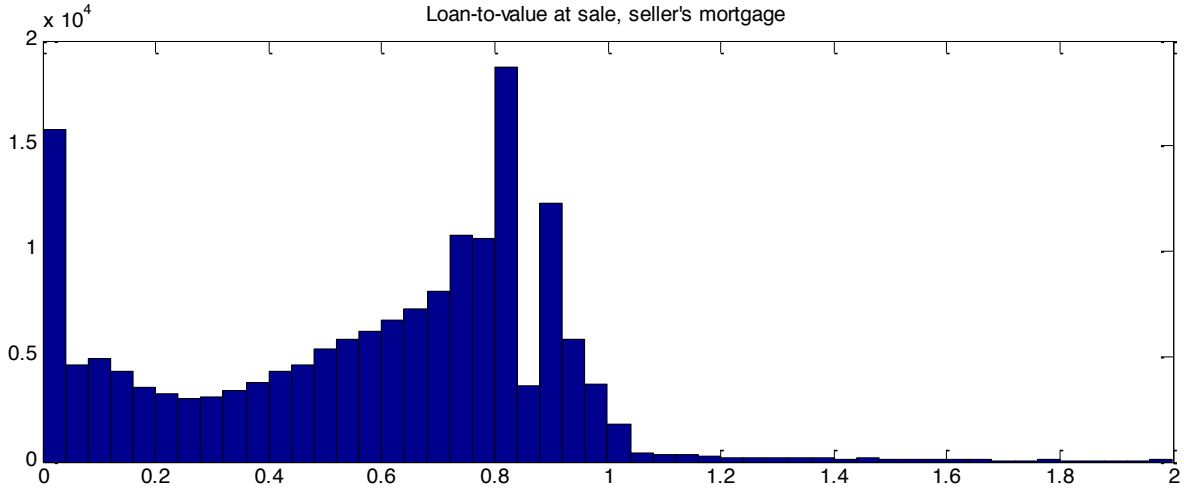


Figure 3: Price Indices without Trade-Selection Process. The figure shows estimated price indices for Alameda County, California, without adjusting for sample selection. The indices are normalized to 100 in the first quarter of 2000. The GLS index is estimated using repeat-sales regressions with weights proportional to the square root of the time between sales. MCMC indices are estimated using the Bayesian procedure described in the text, without including the sales or foreclosure process. Arithmetic indices are calculated with the 1/2 σ^2 adjustment described in the text.

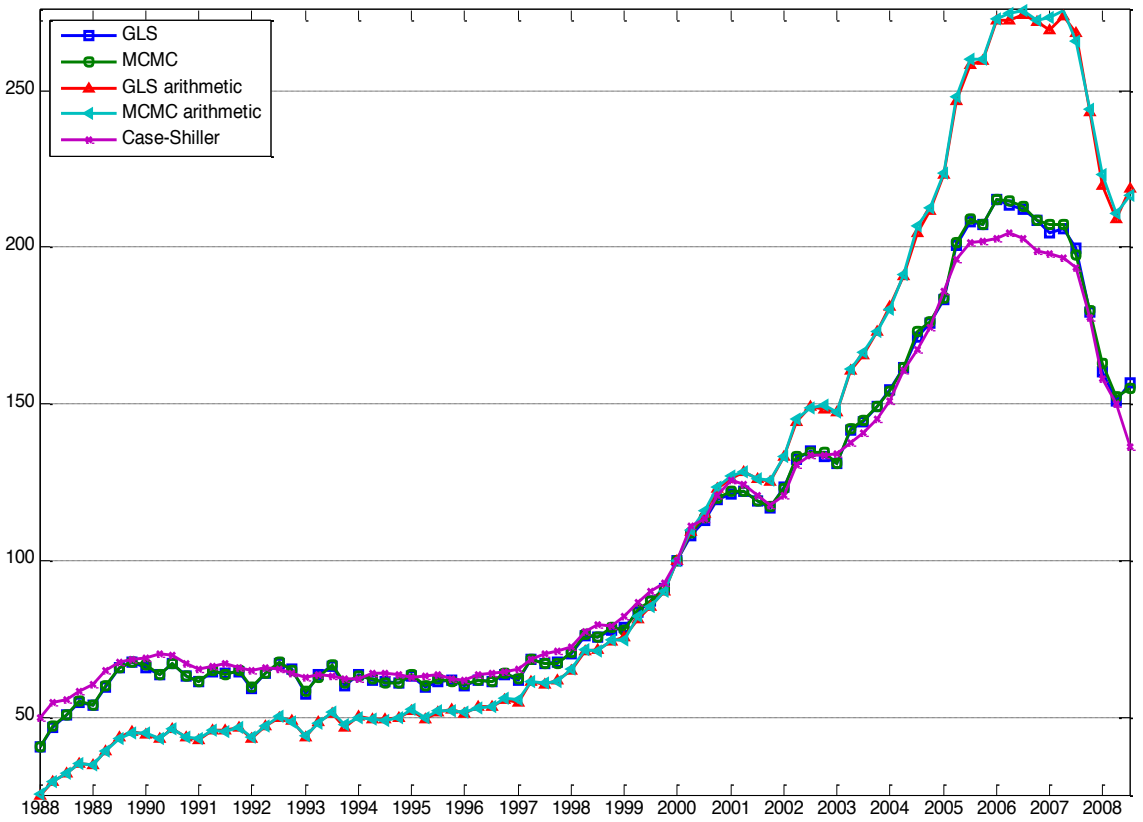


Figure 4: Price Indices with Trade-Selection Process. The figure shows estimated price indices for Alameda County, CA. The indices are normalized to 100 in the first quarter of 2000. All indices are estimated using the Bayesian procedure described in the paper. The baseline index is calculated without adjusting for sample selection. The remaining indices are calculated using four different specifications of the trade process. The bottom plot presents arithmetic indices, calculated with the 1/2 σ2 adjustment discussed in the text. The top graph presents indices without this adjustment.

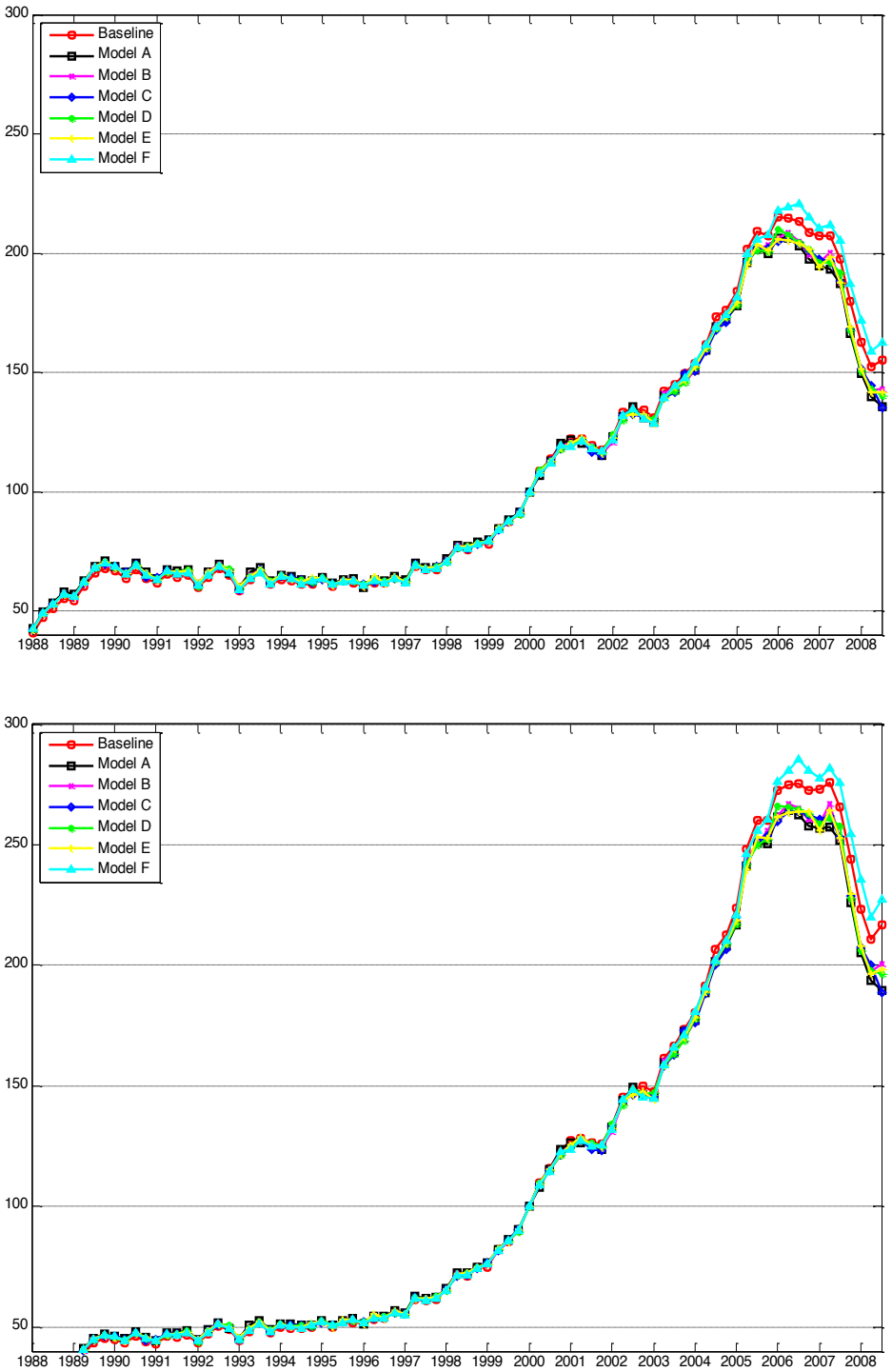
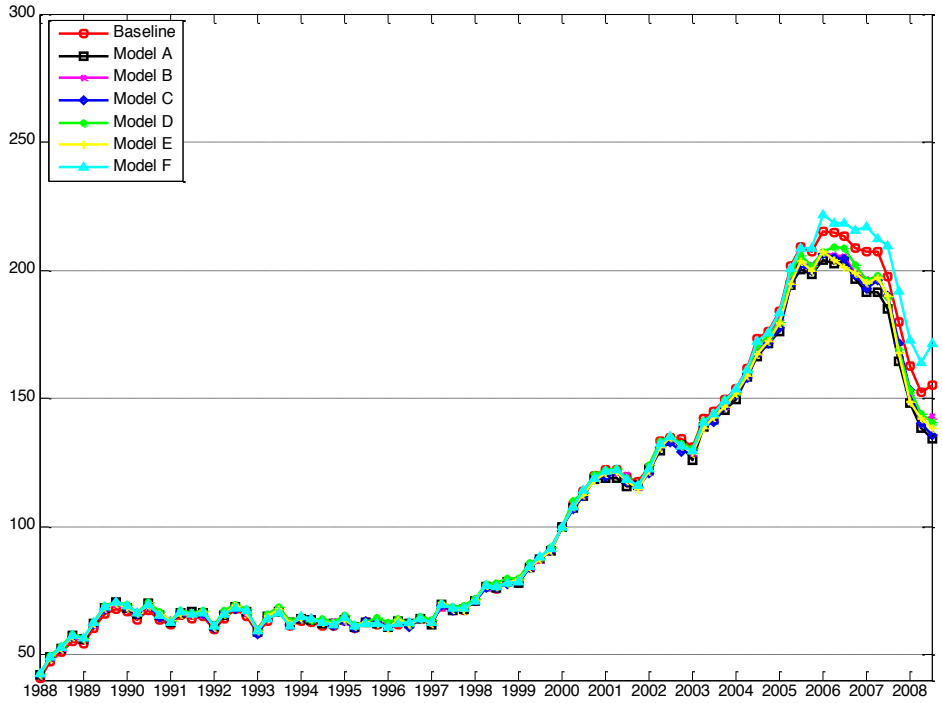


Figure 5: Price Indices with Separate Trade and Foreclosure Processes. The figure shows estimated price indices for Alameda County, CA. The indices are normalized to 100 in the first quarter of 2000. All indices are estimated using the Bayesian procedure described in the paper. The baseline index is calculated without adjusting for sample selection. The remaining indices are calculated using four different specifications of the trade process. The bottom plot presents arithmetic indices, calculated with the 1/2 σ2 adjustment discussed in the text. The top graph presents indices without this adjustment.



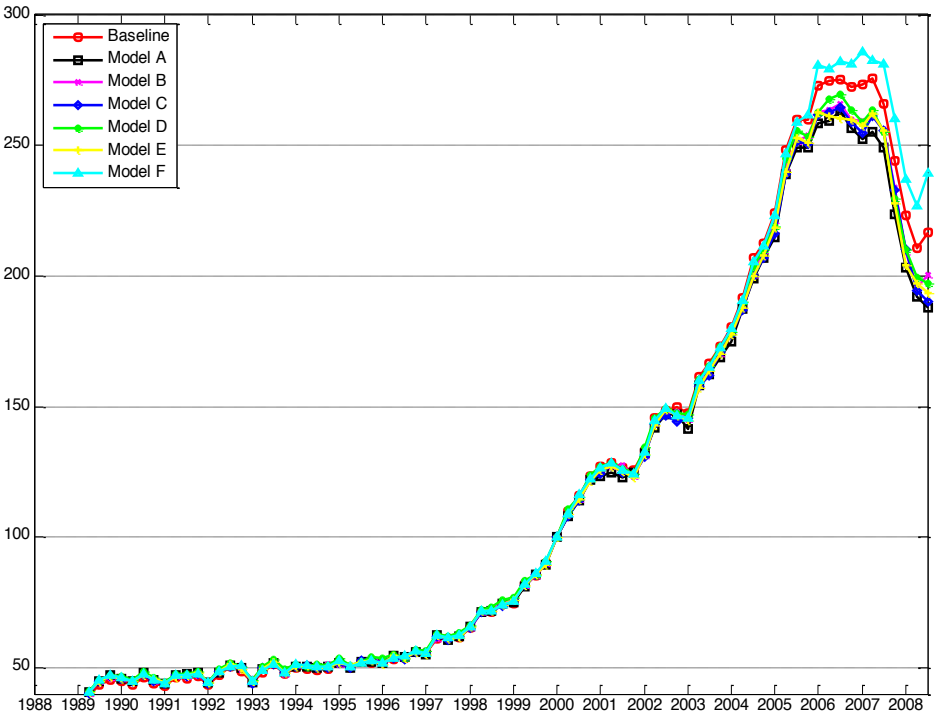


Figure 6: Index Revisions in Calendar Time. Each line represents a quarterly index estimated over a given sample period. All sample periods begin in 1988 and the ending period varies in one-year increments. The top panel presents standard repeat-sales indices. The bottom panel presents MCMC indices estimated with a trade-selection process specified as Model (A). Indices are normalized to one in the first quarter of 1988.

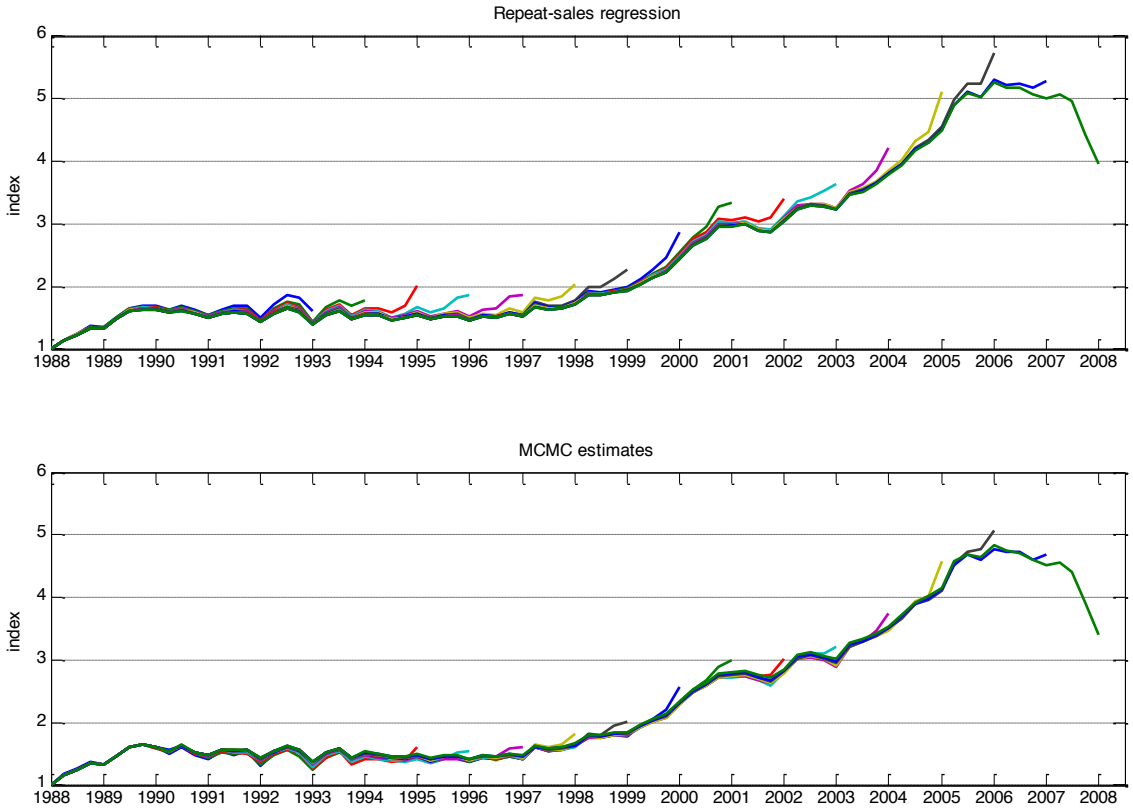


Figure 7: Cumulative Index Revisions in Relative Time. The figures present revisions to the initial index estimate as the sample period is extended beyond the initial period. The horizontal axis gives the number of additional quarters of data used to estimate the index value. The vertical axis is the change in the log index when additional quarters are included relative to the initial index estimate. The solid line gives the median revision, and the dotted lines indicate the 5th and 95th percentiles in the distribution of revisions.

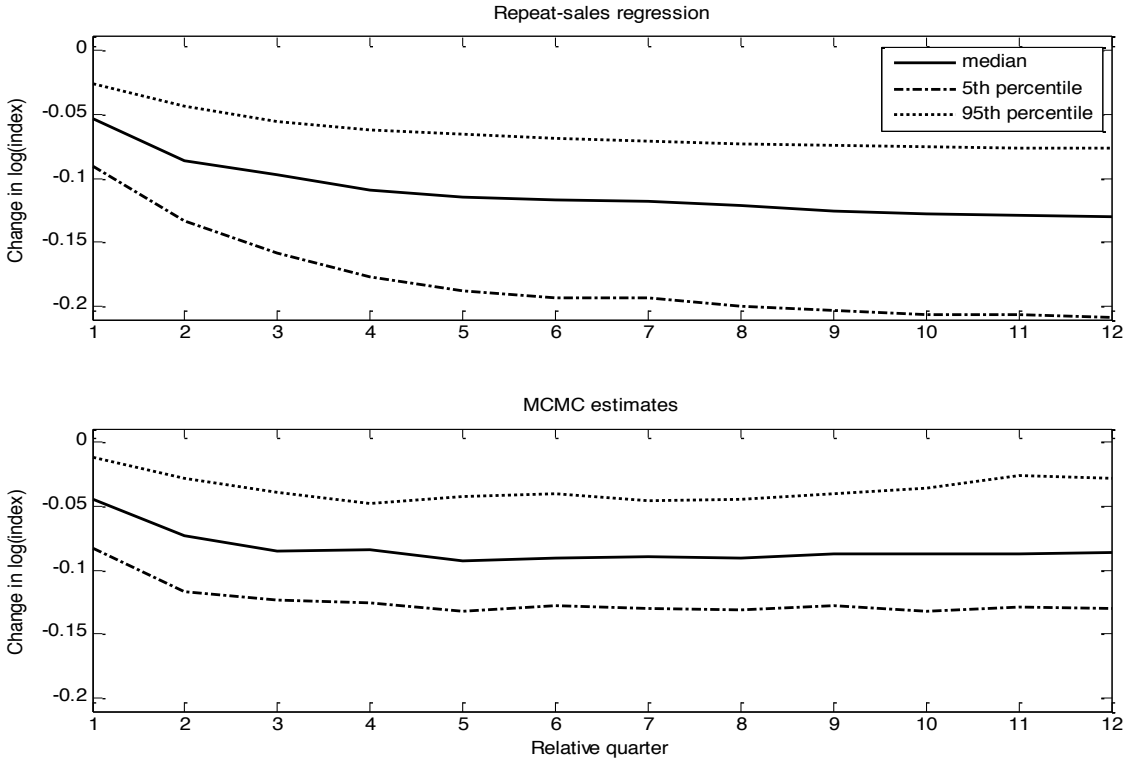


Figure 8: Sellers' LTVs, Foreclosures, and Fractions of Underwater Properties. The figures report the time series of the median of the sellers' LTV at the time of sale (left-hand side scale in each plot), and the fraction of the housing stock that sold (top plot, right-hand scale), foreclosures as a fraction of total sales (middle plot, right-hand scale), and the fraction of properties sold that were underwater at the time of sale (bottom plot, right-hand side).

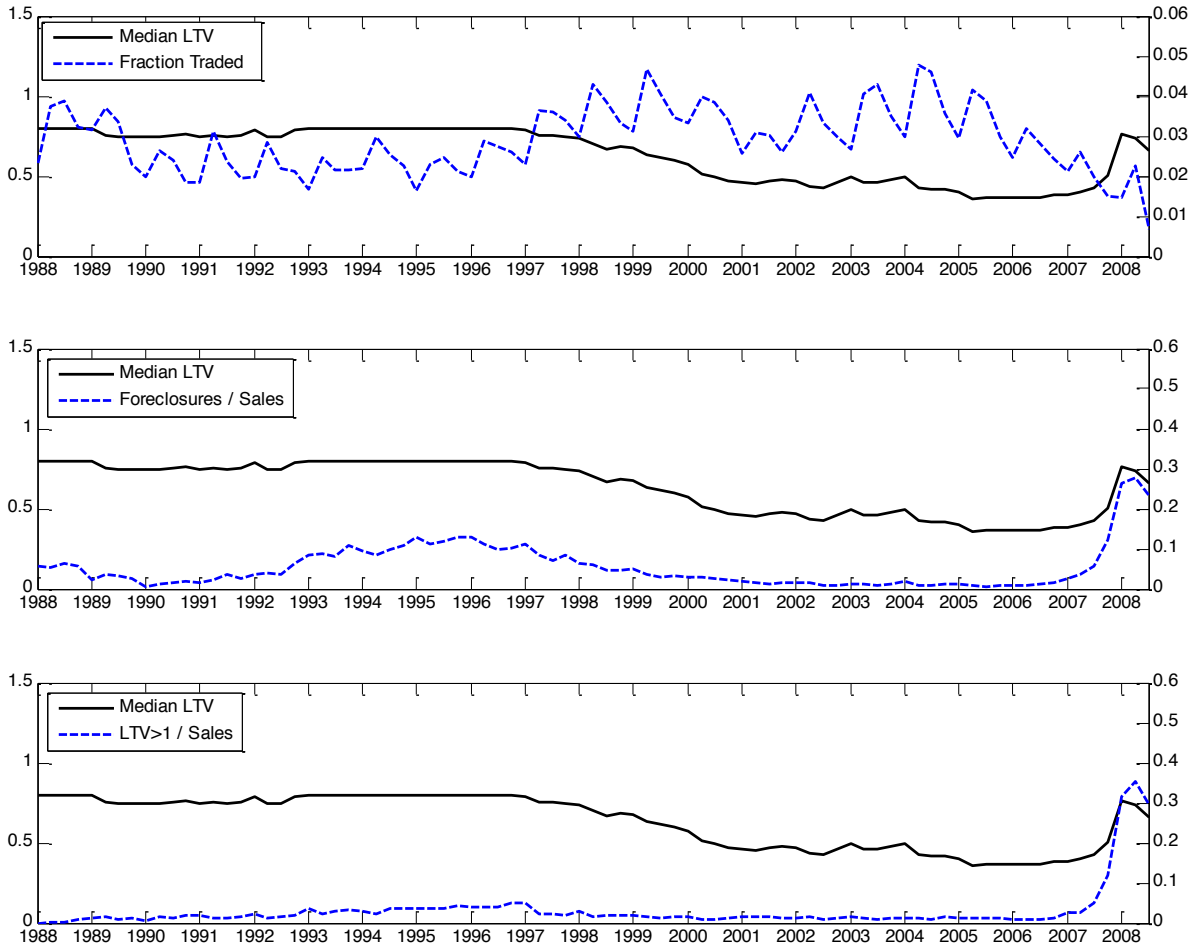
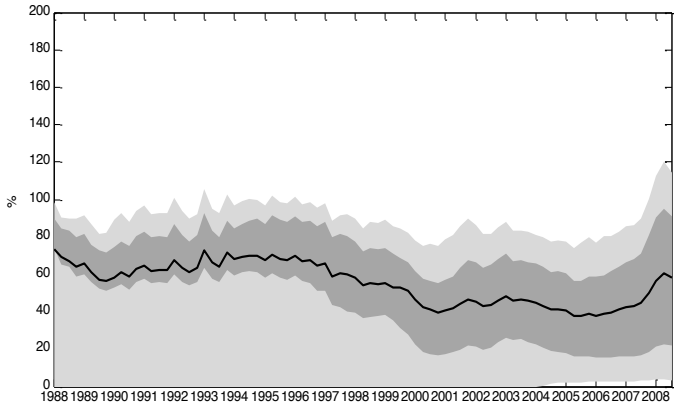
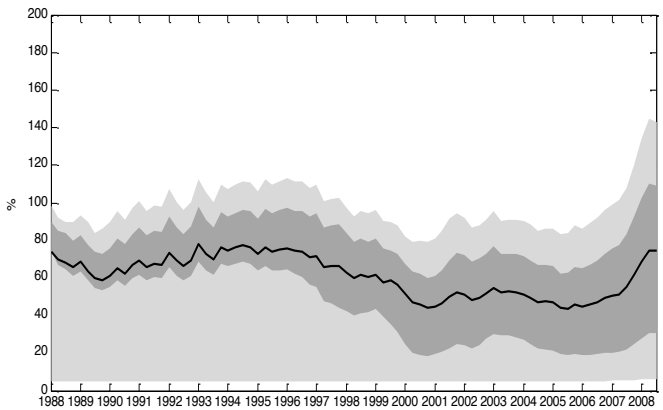


Figure 9: Estimated LTV Percentiles. The plots show estimated time series of the median LTV ratio for the standard repeat-sales (RSR, top-left), the baseline model without selection correction (top-right), and Models A and F of the trade process (see Tables II and IV for model specifications). The light-shaded bands show (5%, 95%) ranges of LTV distributions, and the dark-shaded bands show (25%, 75%) ranges.

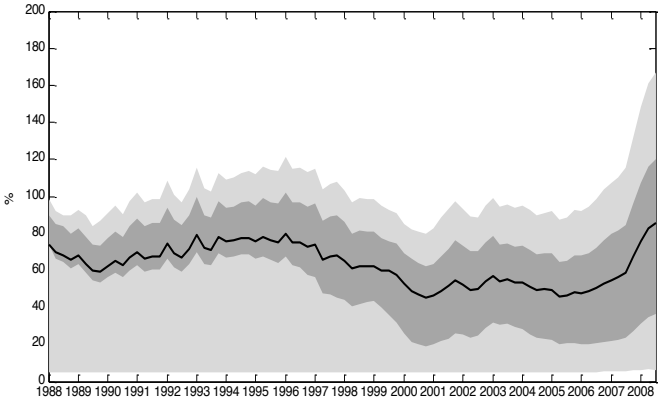
RSR:



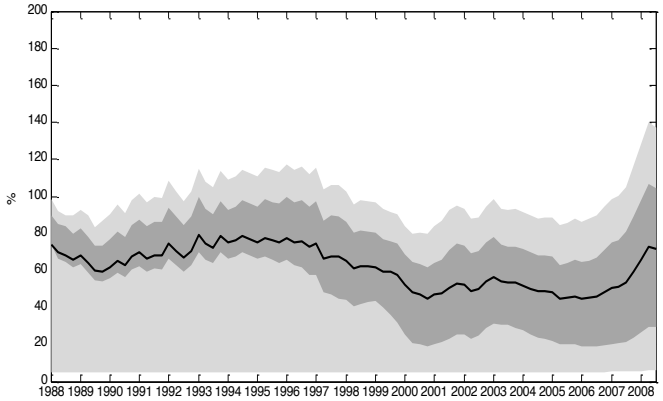
Baseline:



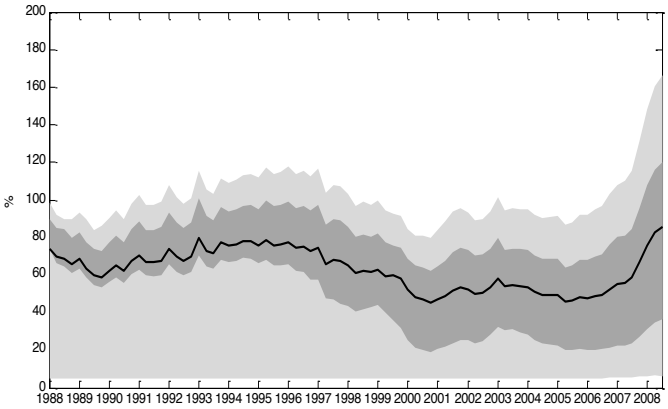
Model A (trade-selection equation):



Model F (trade-selection equation):



Model A (trade and foreclosure selection equation):



Model F (trade and foreclosure selection equation):

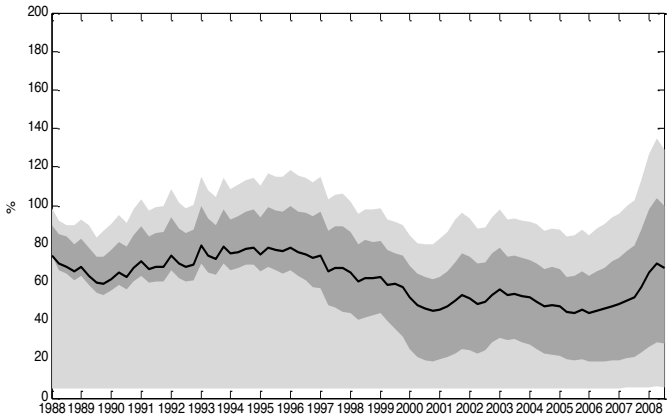


Figure 10: Fraction of Underwater Properties. The figure show time-series plots of the fraction of underwater properties estimated using the standard repeat-sales model (RSR), the baseline model without selection correction, and Models A and F of the trade process (as described in Table II).

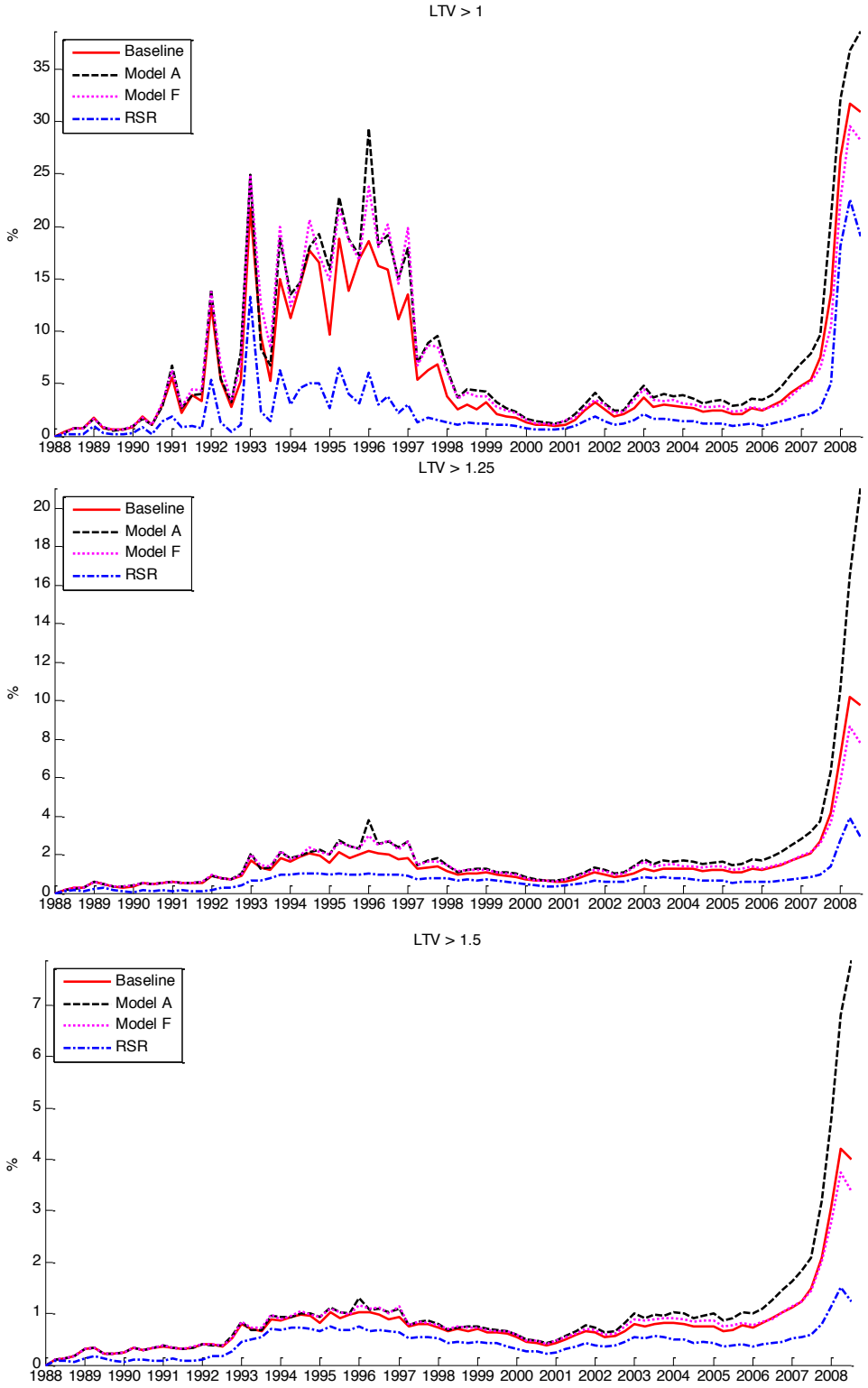


Figure 11: Fraction of Underwater Properties The figure show time-series plots of the fraction of underwater properties estimated using the standard repeat-sales model (RSR), the baseline model without selection correction, and Models A and F of the trade and foreclosure processes (as described in Table IV).

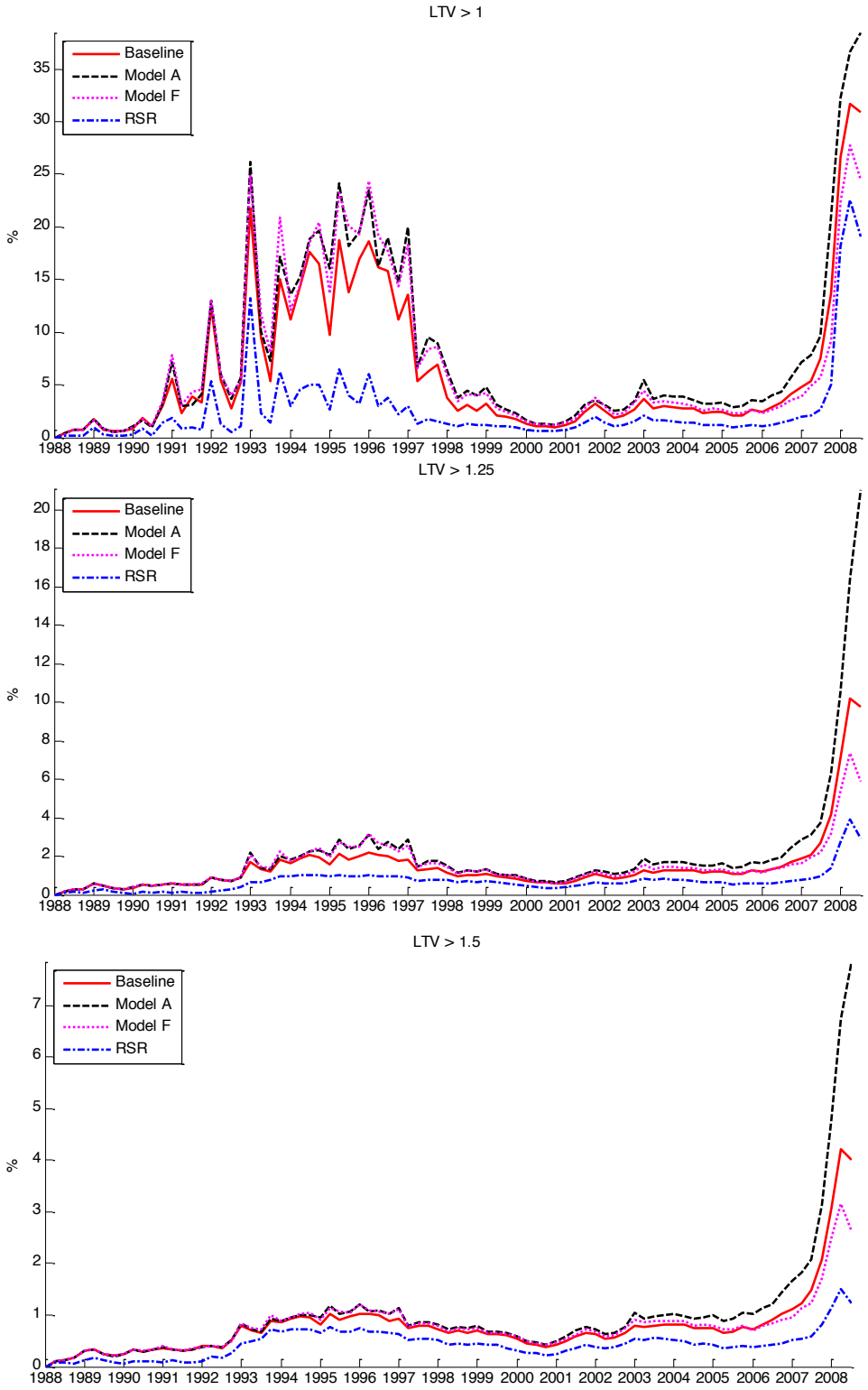
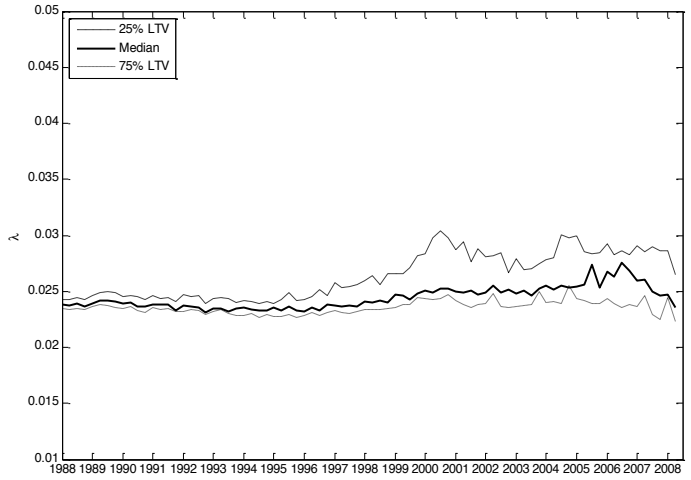
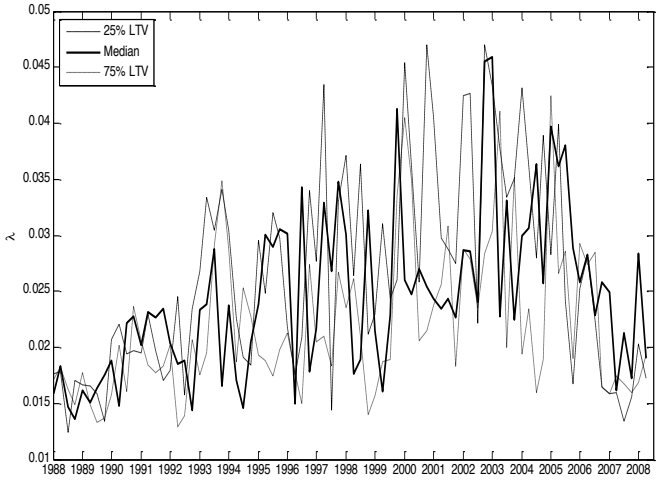


Figure 12: Trade Intensities. The figures show the time series of the constant equivalent trade quarterly intensities (computed as described in the text) at the median LTV ratio and the 25th and 75th percentiles of LTV. The two top plots refer to Models A and F in Table II. The bottom two plots refer to Models A and F in Table IV with separate trade and foreclosure processes.

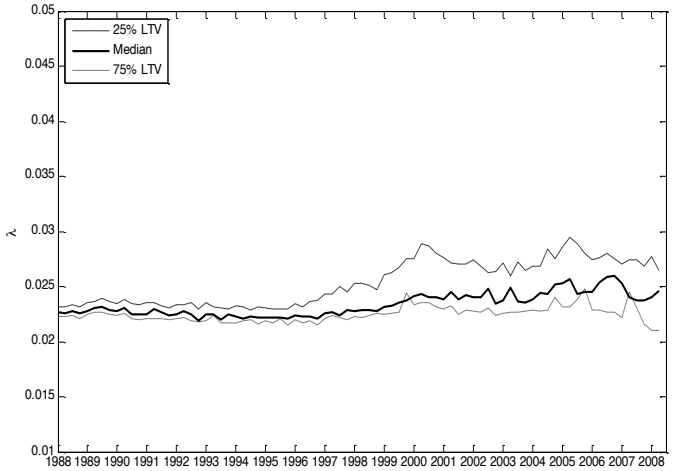
Model A (trade-selection equation):



Model F (trade-selection equation):



Model A (trade and foreclosure selection equation):



Model F (trade and foreclosure selection equation):

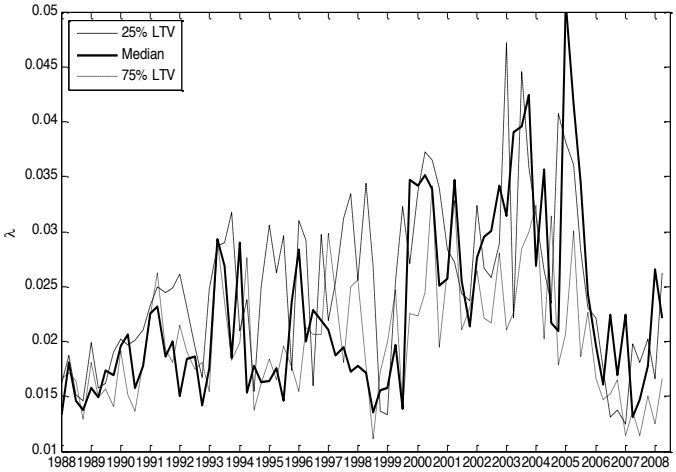
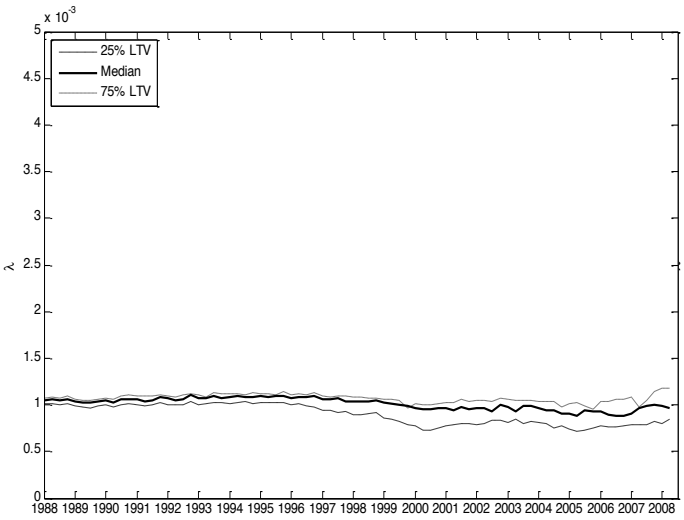


Figure 13: Equivalent Constant Foreclosure Intensities. The figures show the time series of the constant equivalent foreclosure quarterly intensities (computed as described in the text) at the median LTV ratio and the 25th and 75th percentiles of LTV. The two plots refer to Models A and F in Table IV with separate trade and foreclosure processes.

Model A (trade and foreclosure equation):



Model F (trade and foreclosure selection selection equation):

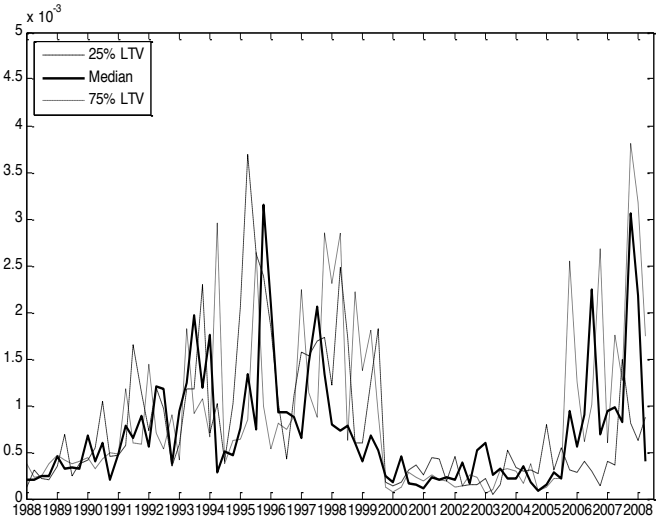


Table I: Summary Statistics. The sample contains single-family residences in Alameda County, California, over the period 1988:Q1 – 2008:Q3. The LTV ratio for the buyer is computed at the time of sale using the buyer's mortgage and the sale price of the home. The seller's LTV uses the outstanding principal of the seller's mortgage, computed as described in the text. Foreclosure rate is the number of foreclosures as a percentage of all properties in the sample.

Panel A: Trade Data

	Mean	Std. Dev.	Min	Q10	Q50	Q90	Max
Number of prop.	68,700						
Number of trades	164,824						
per property	2.40	0.85	1	2	2	3	10
Loan-to-value ratio							
buyer	0.71	0.26	0.00	0.24	0.80	0.95	1.00
seller	0.58	0.48	0.00	0.05	0.65	0.90	68.79
Sale-to-sale return (%)							
arithmetic	64.44	79.34	-86.40	-4.90	42.26	167.52	500.00
log	40.01	44.38	-199.51	-4.65	35.54	98.25	179.18
Time between sales (years)	5.15	3.71	0.25	1.00	4.25	10.50	20.50
Foreclosure rate (%)	0.13						

Panel B: Property Characteristics

	Mean	Std. Dev.	Min	Q10	Q50	Q90	Max
Acres	1.34	0.68	0.03	0.62	1.24	2.12	5.00
Space (000s sqft.)	1.66	0.64	0.34	1.00	1.52	2.52	7.53
Bedrooms (#)	3.15	0.88	1	2	3	4	9
Bathrooms (#)	2.17	0.85	1	1	2	3	5
Total rooms (#)	6.50	1.58	2	5	6	9	15
Basement	0.03						
Garage	0.90						
Fireplace	0.33						
Pool	0.06						
Stories	1.31	0.39	1	1	1	2	3
Construction year	1962.95	25.89	1901	1924	1965	1996	2006

Table II: Trade-Selection Process: *LTV* is the natural logarithm of the LTV ratio. *Time* indicates time since the previous sale (in years). *ΔMortgage rate* is the change in the 30-year mortgage rate since the inception of the loan. *Square footage* is the property's size in thousands of square feet. *Sigma* is the annualized standard deviation of the error term in the observation (price index) equation. Posterior standard deviations are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	A	B	C	D	E	F
LTV	-0.0598***	-0.0401***	-0.0448***	-0.0426***	-0.0406***	
	(0.0012)	(0.0017)	(0.0015)	(0.0014)	(0.0051)	
LTV 1988-1993						-0.0789***
						(0.0041)
LTV 1994-1999						-0.0504***
						(0.0026)
LTV 2000-2005						-0.0457***
						(0.0020)
LTV 2006-2008						0.0389***
						(0.0033)
LTV x					0.0130***	
Square Footage					(0.0024)	
LTV x					-0.0006***	
Age (years)					(0.0001)	
Time (years)		0.0505***	0.0381***	0.0511***	0.0519***	0.0413***
		(0.0010)	(0.0013)	(0.0011)	(0.0011)	(0.0012)
Time Squared		-0.0022***	-0.0021***	-0.0022***	-0.0023***	-0.0020***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
ΔMortgage rate			-5.4128***			-3.6849***
			(0.1463)			(0.1939)
Square footage				-0.0902***	-0.0671***	-0.0809***
				(0.0040)	(0.0061)	(0.0047)
Age (years)				-0.0008***	-0.0013***	-0.0006***
				(0.0001)	(0.0001)	(0.0001)
Intercept	-2.0073***	-2.1199***	-2.1132***	-2.0573***	-2.0471***	
	(0.0017)	(0.0034)	(0.0046)	(0.0055)	(0.0062)	
Intercept 1988-93						-2.1061***
						(0.0075)
Intercept 1994-99						-2.0671***
						(0.0065)
Intercept 2000-05						-2.0036***
						(0.0059)
Intercept 2006-08						-2.0664***
						(0.0079)
Seasonal Dummies	No	Yes	Yes	Yes	Yes	Yes
Sigma	0.2814***	0.2816***	0.2813***	0.2812***	0.2815***	0.2810***
	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0005)	(0.0005)

Table III: LTV Coefficient Across Samples. The repeat-sales sample is the sample that is described in Table I and that is used in all other tables and figures. The single-sales sample includes all properties that had at least one trade during the sample period. The no foreclosures sample is the repeat-sales sample without properties that experienced a foreclosure during the sample period. *LTV* is the natural logarithm of the LTV. *Time* indicates the time since the previous sale (in years). *Sigma* is the annualized standard deviation of the error term in the observation (price index) equation. Posterior standard deviations are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Repeat Sales	Single Sales	No Foreclosures
LTV	-0.0598***	-0.0203***	-0.0620***
	(0.0012)	(0.0014)	(0.0017)
Intercept	-2.0073***	-2.2122***	-2.0125***
	(0.0017)	(0.0017)	(0.0021)
Seasonal Dummies	No	No	No
Sigma	0.2814***	0.2812***	0.2695***
	(0.0004)	(0.0005)	(0.0005)
# Properties	68,700	142,794	41,983

Table IV: Trade and Foreclosure Processes: *LTV* is the natural logarithm of the LTV ratio. *Time* indicates time since the previous sale (in years). *ΔMortgage rate* is the change in the 30-year mortgage rate since the inception of the loan. *Square footage* is the property's size in thousands of square feet. *Sigma* is the annualized standard deviation of the error term in the observation (price index) equation. Posterior standard deviations are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A: Trade Selection Process

	A	B	C	D	E	F
LTV	-0.0656***	-0.0455***	-0.0513***	-0.0486***	-0.0443***	
	(0.0017)	(0.0018)	(0.0017)	(0.0019)	(0.0053)	
LTV 1988-1993						-0.0810***
						(0.0043)
LTV 1994-1999						-0.0500***
						(0.0024)
LTV 2000-2005						-0.0425***
						(0.0027)
LTV 2006-2008						0.0122***
						(0.0041)
LTV *					0.0148***	
Square Footage					(0.0026)	
LTV *					-0.0007***	
Age (years)					(0.0001)	
Time (years)		0.0482***	0.0358***	0.0487***	0.0493***	0.0402***
		(0.0012)	(0.0011)	(0.0012)	(0.0012)	(0.0012)
Time Squared		-0.0020***	-0.0020***	-0.0020***	-0.0020***	-0.0018***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
ΔMortgage rate			-5.3842***			-3.1686***
			(0.1537)			(0.1919)
Square footage				-0.0835***	-0.0558***	-0.0706***
				(0.0047)	(0.0068)	(0.0046)
Age (years)				-0.0010***	-0.0016***	-0.0008***
				(0.0001)	(0.0001)	(0.0001)
Intercept	-2.0298***	-2.1325***	-2.1278***	-2.0666***	-2.0522***	
	(0.0025)	(0.0043)	(0.0040)	(0.0059)	(0.0067)	
Intercept 1988-93						-2.1075***
						(0.0073)
Intercept 1994-99						-2.0839***
						(0.0061)
Intercept 2000-05						-1.9943***
						(0.0062)
Intercept 2006-08						-2.1233***
						(0.0081)
Seasonal Dummies	No	Yes	Yes	Yes	Yes	Yes
Sigma	0.2815***	0.2816***	0.2816***	0.2813***	0.2815***	0.2806***
	(0.0004)	(0.0004)	(0.0004)	(0.0007)	(0.0004)	(0.0004)

Panel B: Foreclosure Selection Process

	A	B	C	D	E	F
LTV	0.0728***	0.0800***	0.0788***	0.0714***	0.0176	
	(0.0038)	(0.0040)	(0.0038)	(0.0031)	(0.0145)	
LTV 1988-1993						0.0019
						(0.0150)
LTV 1994-1999						-0.0283***
						(0.0067)
LTV 2000-2005						-0.1292***
						(0.0041)
LTV 2006-2008						0.4018***
						(0.0084)
LTV *					0.0155*	
Square Footage					(0.0065)	
LTV *					0.0008***	
Age (years)					(0.0002)	
Time (years)		0.0877***	0.0750***	0.0877***	0.0844***	0.0516***
		(0.0038)	(0.0038)	(0.0038)	(0.0064)	(0.0026)
Time Squared		-0.0061***	-0.0057***	-0.0063***	-0.0061***	-0.0050***
		(0.0003)	(0.0002)	(0.0003)	(0.0004)	(0.0002)
ΔMortgage rate			-3.5461***			-8.7778***
			(0.3114)			(0.5034)
Square footage				-0.1863***	-0.1841***	-0.1805***
				(0.0131)	(0.0181)	(0.0120)
Age (years)				0.0029***	0.0036***	0.0028***
				(0.0002)	(0.0002)	(0.0001)
Intercept	-3.0504***	-3.3177***	-3.2843***	-3.3598***	-3.3867***	
	(0.0055)	(0.0140)	(0.0162)	(0.0232)	(0.0193)	
Intercept 1988-93						-3.4580***
						(0.0219)
Intercept 1994-99						-3.2079***
						(0.0094)
Intercept 2000-05						-3.8252***
						(0.0112)
Intercept 2006-08						-2.9094***
						(0.0155)
Seasonal Dummies	No	Yes	Yes	Yes	Yes	Yes

Table V: Marginal Intensities: Panel A shows the change in the equivalent constant sale intensity from a one-standard deviation increase in LTV, time since the last sale and its squared value, 30-year mortgage rates, house square footage, and age. Panel B contains corresponding figures for the foreclosure intensity. Results are based on the models in Table IV. Posterior standard deviations are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A: Change in Trade Intensity ($x10^{-3}$)

	A	B	C	D	E	F
LTV	-4.8697***	-3.8320***	-4.4917***	-3.3861***	-3.4406***	
	(0.1143)	(0.1471)	(0.1365)	(0.1316)	(0.1253)	
LTV 1988-1993						-5.6159***
						(0.2933)
LTV 1994-1999						-3.4669***
						(0.1663)
LTV 2000-2005						-2.9507***
						(0.1874)
LTV 2006-2008						0.8453***
						(0.2849)
Time (years)		7.0008***	4.4847***	5.8855***	6.3980***	4.6663***
		(0.1400)	(0.1553)	(0.1300)	(0.1602)	(0.1535)
ΔMortgage rate			-6.1567***			-2.8592***
			(0.1865)			(0.1594)
Square footage				-2.6833***	-2.0364***	-2.2606***
				(0.1269)	(0.1900)	(0.1281)
Age (years)				-1.2399***	-2.0688***	-1.0443***
				(0.0737)	(0.1080)	(0.0697)

Panel B: Change in Foreclosure Intensity ($x10^{-3}$)

	A	B	C	D	E	F
LTV	0.3658***	0.3949***	0.4280***	0.1708***	0.1599***	
	(0.0230)	(0.0156)	(0.0338)	(0.0091)	(0.0173)	
LTV 1988-1993						0.0047
						(0.0265)
LTV 1994-1999						-0.0481***
						(0.0099)
LTV 2000-2005						-0.2218***
						(0.0171)
LTV 2006-2008						0.6901***
						(0.0531)
Time (years)		0.4789***	0.3916***	0.2176***	0.2046***	0.0411***
		(0.0284)	(0.0345)	(0.0139)	(0.0235)	(0.0093)
ΔMortgage rate			-0.2514***			-0.1963***
			(0.0273)			(0.0116)
Square footage				-0.2053***	-0.2016***	-0.1427***
				(0.0108)	(0.0052)	(0.0092)
Age (years)				0.1277***	0.1503***	0.0895***
				(0.0108)	(0.0176)	(0.0100)