A New Dynamic House-Price Index for Mortgage Valuation and Stress Testing

Richard Stanton Nancy Wallace

Haas School of Business, U.C. Berkeley

CSRA Winter meeting
December 2015

Introduction

- Unlike stocks, houses trade very infrequently.
- So we construct **house price indices** to tell us about price dynamics of the "average" house. Important both historically and for forecasting.
  - Mortgage valuation.
  - Bank stress tests.
  - Testing for housing market efficiency.
  - Analyzing the price formation process of housing markets.
  - Automated appraisals, e.g, Zillow, property taxes.
  - Measure of health of the housing market/economy.
  - Valuation of housing-market derivatives.
- If we use a poor index, our decisions/valuations will be suspect.
- **This paper:** We point out some serious problems with existing indices (and data) and develop a better approach.

Repeat-sales indices

- Calhoun (1996) describes methodology used to construct FHFA (formerly OFHEO) index.
  - Methodologies for different indexes differ slightly.
- Write price of house i at time t as

$$p_{it} = \log(P_{it}) = \beta_t + H_{it} + N_{it},$$

  where
  - P_{it} = house price
  - H_{it} = Gaussian random walk
  - N_{it} = white noise
- Total return for a house selling in periods s and t is

$$\begin{aligned} y_i &= p_{it} - p_{is} \\ &= (\beta_t + H_{it} + N_{it}) - (\beta_s + H_{is} + N_{is}). \end{aligned}$$

Repeat-sales indices

- We can write this as

$$y_i = \sum_{\tau=0}^{T} (\beta_\tau + H_{i\tau} + N_{i\tau}) D_{i\tau} = \sum_{\tau=0}^{T} \beta_\tau D_{i\tau} + \epsilon_i,$$

where

$$D_{i\tau} = \begin{cases} +1 & \text{if house } i \text{ sold for a second time at time } \tau, \\ -1 & \text{if house } i \text{ sold for the first time at time } \tau, \\ 0 & \text{otherwise.} \end{cases}$$

$$\epsilon_i = (H_{it} - H_{is}) + (N_{it} - N_{is}), \qquad \text{so}$$

$$\text{var}(\epsilon_i) = A(t-s) + B(t-s)^2 + C.$$

- **Estimate index levels, β_τ, by regressing y_i on D_i using OLS (GLS).**

Problems with repeat-sales indices 1

Most houses don't sell very often

- A house is only included if it sells more than once, so we throw out a huge fraction of house sales.
- In San Francisco between 2003–2012,
  - 24,342 houses sold at least once.
  - 20,778 (85.4%) sold only once, so are discarded.
  - Index considers only 3,564 houses (14.6%)
- In Los Angeles between 2003–2012,
  - 236,406 houses sold at least once.
  - 194,375 (82.2%) sold only once, so are discarded.
  - Index considers only 42,031 houses (17.8%)

Problems with repeat-sales indices 2

Changes in property characteristics

- In San Francisco, 2502 Leavenworth sold in 2003 and 2008.
  - In 2003:
    - Square footage = 1,752
    - Total rooms = 5
    - Bathrooms = 1
    - Price = $1,503,000
  - In 2008:
    - Square footage = **2,913**
    - Total rooms = **8**
    - Bathrooms = **3**
    - Price = **$5,500,000**
- Changes in size (or quality) are wrongly counted as “returns”

Problems with repeat-sales indices 3

The index is not a house price!

- For mortgage valuation, stress testing, etc., we need the distribution of **future** house prices.
- In repeat-sales methodology, each period's index growth is a **constant**.
  - No specification of inter-period dynamics.
- Even if we assume house prices follow (say) geometric Brownian motion, we can't just use the volatility of the index.
  - Index levels are **estimated** (though standard errors are never shown).
  - Construction of index automatically induces smoothing.
  - Volatility of index will underestimate volatility of individual house prices.
  - Even ignoring this, index is a (somewhat) diversified portfolio, not an individual house.
- We also can't use index properties to test for (e.g.) serial correlation in house prices.
  - Estimation procedure mechanically induces serial correlation in index.
    - Even when there is none in individual house prices.

Data Issues

- Until recently, it was not possible to obtain accurate home characteristics in the U.S.
- Recently, some vendors starting collecting (and selling) home characteristic data from county recorders' offices.
- 7 years ago we started thinking about this project!
- Negotiated hard with CoreLogic to get price and characteristic data.
  - Got them down to $40,000 for a license to use data on San Francisco and Alameda for one year!
- It was completely useless!
  - Every year, they **overwrite** the property characteristics with new data!
  - So we only get a **single snapshot** (taken at an unknown time).

A New Data Set

- In response to our criticisms of existing data sets, DataQuick went back to their backup tapes and put together **DataQuick housing stock data** (released October 2013)
  - Annual snapshot of property-specific characteristics for stock of housing, including remodels, 2003–2012.
- We merge this with **DataQuick historical transaction data**
  - (Static) home characteristic data + parcel-specific transaction histories.
- Merging these data sets, we create a sample of the stock of existing single-family residential houses, including:
  - Time-varying characteristics for each house.
  - All arms-length transactions including foreclosure auction sales.
  - Six large urban counties in California (Alameda, Contra Costa, Los Angeles, San Diego, San Francisco, and Santa Clara).
- Uncertain future: DataQuick was bought by CoreLogic in March 2014.

New index

- We want to construct an index that overcomes all of the problems with prior indices.
- In particular,
  - Explicit consideration of dynamics.
  - Dependence on property characteristics.
  - Dependence on macroeconomic variables.
  - Include all sales, not just repeats.
  - Allow for unobserved heterogeneity across properties.

New Index

Write log-price, $y_{i,t}$, as

$$
\begin{aligned}
y_{i,t} &= \overbrace{A_b x_{i,t} + B_b \xi_t + \alpha_{b,t}}^{\text{house price index}} + \mu_i + \epsilon_{i,t}, \\
&= \boldsymbol{X}_t \beta_b + \alpha_{b,t} + \mu_i + \epsilon_{i,t}, \\
\alpha_{b,t} &= \rho \alpha_{b,t-1} + \eta_t, \qquad \text{where} \\
\epsilon_{i,t} &\sim \text{i.i.d. } N(0, \sigma_\epsilon^2), \\
\mu_i &\sim \text{i.i.d. } N(0, \sigma_\mu^2), \\
\eta_{i,t} &\sim \text{i.i.d. } N(0, \sigma_\eta^2).
\end{aligned}
$$

- $x_{i,t}$ = home hedonics (beds, size, lot size).
- ξ_t = macro fundamentals (interest rates, population, unemployment).
- μ_i = house-specific random effect (due to unobservable differences).
- $\alpha_{b,t}$ = unexplained (and unobserved) portion of the index.

Estimation

- Model can be written as a standard, linear state-space model if we augment the state, s_t, to include μ_i:

$$s_t = \begin{pmatrix} \alpha_{b,t} \\ \beta_{1,t} \\ \beta_{2,t} \\ \vdots \\ \beta_{k,t} \\ \mu_{1,t} \\ \mu_{2,t} \\ \vdots \\ \mu_{I,t} \end{pmatrix} \equiv \begin{pmatrix} \alpha_{b,t} \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix} = \begin{pmatrix} \alpha + \rho\alpha_{b,t-1} + \eta_t \\ \beta_{1,t-1} \\ \beta_{2,t-1} \\ \vdots \\ \beta_{k,t-1} \\ \mu_{1,t-1} \\ \mu_{2,t-1} \\ \vdots \\ \mu_{I,t-1} \end{pmatrix}.$$

Estimation

- In principle, we could estimate the model's parameters by maximizing the likelihood function for this state space model using standard Kalman Filter (Kalman and Bucy, 1961).
- We don't do it this way.
  - Dimensionality of problem is very high.
  - Lots of missing data.
  - Want to be able to extend to even more general settings.
    - E.g., more general distributional assumptions.
- Instead we use **Markov Chain Monte Carlo (MCMC)** Bayesian methods (e.g., Johannes and Polson, 2009).
  - Details in paper.

Table 1: Summary Statistics

	Mean	Standard Deviation	Mean	Standard Deviation
	Alameda		**Contra Costa**	
Structure Square Footage	1,743.05	900.55	1,828.79	805.37
Number of Bathrooms	1.99	0.89	2.19	0.79
Number of Bedrooms	3.17	0.92	3.40	0.85
Lot Size Square Footage	6,987.93	13,076.47	8,807.51	11,201.89
Year Built	1958.68	28.78	1973.60	23.00
Price ($)	550,299.28	361,461.07	470,422.4	395,990.83
Arms-length Transactions	123,554		153,701	
	Los Angeles		**San Diego**	
Structure Square Footage	1,623.30	800.84	1,783.38	949.65
Number of Bathrooms	1.97	0.91	2.12	0.84
Number of Bedrooms	3.06	0.88	3.25	0.86
Lot Size Square Footage	8,404.70	9,487.27	22,927.14	53,937.76
Year Built	1950.49	15.66	1966.89	17.66
Price ($)	595,518.31	724,380.42	548,628.56	538,721.78
Arms-length Transactions	286,407		103,749	
	San Francisco		**Santa Clara**	
Structure Square Footage	1,570.41	821.86	1,810.72	788.61
Number of Bathrooms	1.54	0.83	2.20	0.78
Number of Bedrooms	N.A.	N.A.	3.48	0.87
Lot Size Square Footage	2,875.63	1,186.66	9,909.73	22,662.21
Year Built	1931.91	21.18	1967.30	21.96
Price ($)	896,693.78	968,338.08	752,923.12	664,236.88
Arms-length Transactions	27,614		113,776	

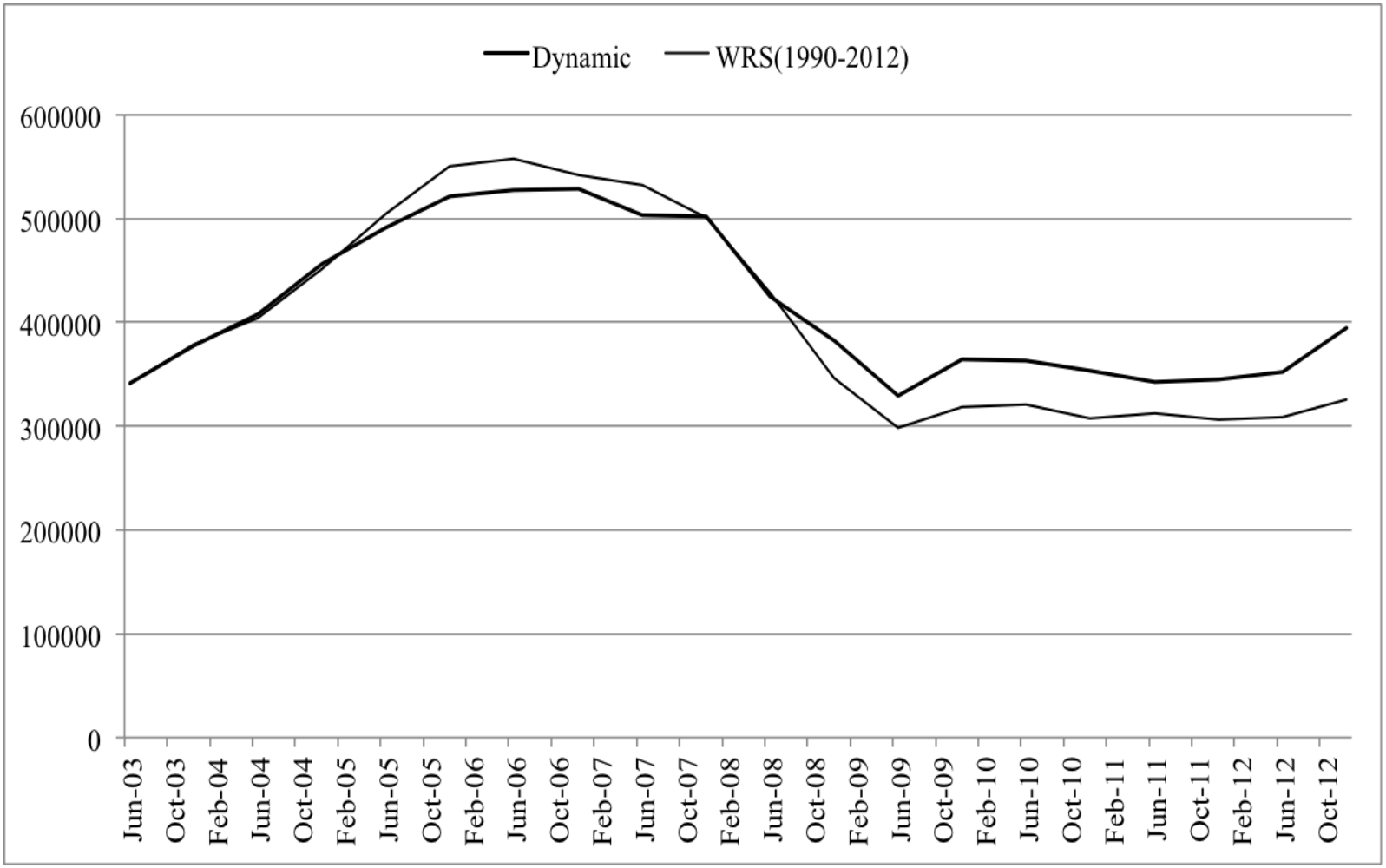
Results 1

	Alameda			
	Weighted Average Across Zip Codes		Overall County	
	Coeff. Est.	Std. Dev.	Coeff. Est.	Std. Err.
σ_ϵ	0.41053	0.06127	0.42421	0.00298
σ_η	0.10106	0.02167	0.07995	0.01866
ρ	0.96585	0.01655	0.95484	0.04950
Number of Bedrooms	0.01005	0.0426	0.02084	0.00280
Lotsize	0.00001	0.00001	0.00000	0.00000
Square Footage	0.00032	0.0001	0.00041	0.00000
Ten year Treasury	0.04793	0.03248	0.04447	0.00324
County Population	0.32758	0.04192	0.33082	0.11955
Unemployment	-0.05437	0.01899	-0.06058	0.01897
Number of Observations	zips = 47		N = 76,789	
	San Francisco			
	Weighted Average Across Zip Codes		Overall County	
	Coeff. Est.	Std. Dev	Coeff. Est.	Std. Err.
σ_ϵ	0.43409	0.06178	0.39810	0.00545
σ_η	0.08947	0.0259	0.04993	0.01128
ρ	0.95956	0.043	0.99909	0.01116
Number of bathrooms	0.02521	0.04658	0.03166	0.00621
Lotsize	0.00004	0.00004	0.00005	0.00000
Square Footage	0.00021	0.00008	0.00037	0.00001
Ten year Treasury	0.03856	0.03731	0.00934	0.00190
County Population	0.30884	0.07855	0.10660	0.07545
Unemployment	-0.03849	0.01454	-0.03484	0.01164
Number of Observations	zips = 20		N = 24,339	
	Contra Costa			
	Weighted Average Across Zip Codes		Overall County	
	Aver. Coeff. Est	Std Dev	Coeff. Est.	Std. Err.
σ_ϵ	0.31373	0.0726	0.33217	0.00402
σ_η	0.11917	0.06818	0.18698	0.07955
ρ	0.95874	0.05798	0.61467	0.30245
Number of Bedrooms	0.00569	0.03921	0.01084	0.00454
Lotsize	0.00001	0.00001	0.00000	0.00000
Square Footage	0.00032	0.00009	0.00035	0.00001
Ten year Treasury	0.06475	0.03933	0.23091	0.15107
County Population	0.29079	0.08147	0.37986	0.07306
Unemployment	-0.05603	0.02287	-0.10825	0.05490
Number of Observations	zips = 39		N = 87,916	

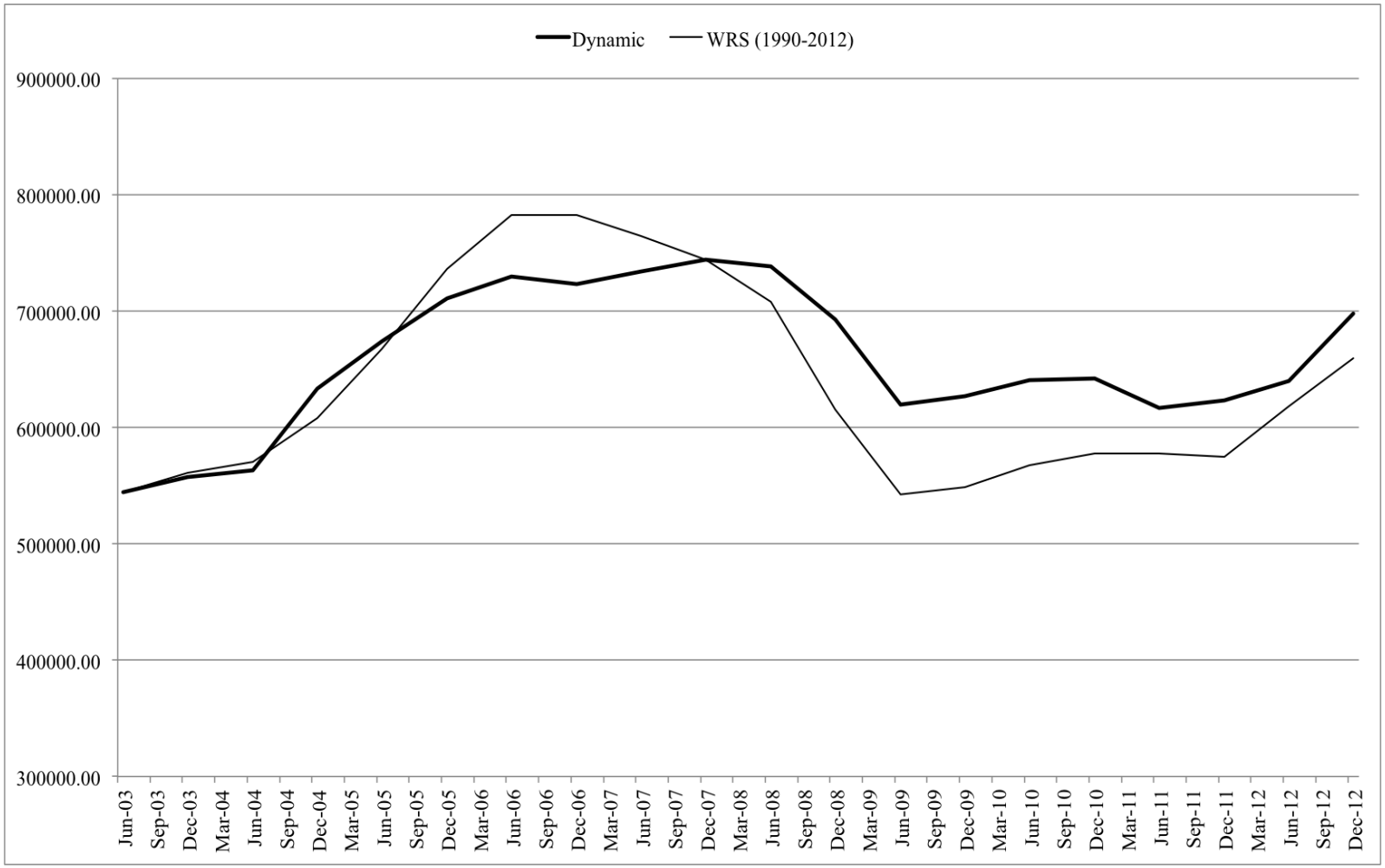
Results 2

	Santa Clara			
	Weighted Average Across Zip Codes		Overall County	
	Coeff. Est.	Std. Dev.	Coeff. Est.	Std. Err.
σ_ϵ	0.44920	0.07789	0.49025	0.00961
σ_η	0.11119	0.06166	0.07994	0.02108
ρ	0.93942	0.10051	0.97310	0.04232
Number of Bedrooms	-0.02979	0.05298	0.16092	0.00835
Lotsize	0.00000	0.00001	0.00000	0.00000
Square Footage	0.00033	0.00009	0.00053	0.00001
Ten year Treasury	0.05341	0.05268	0.04536	0.03324
County Population	0.33597	0.06521	0.32486	0.12246
Unemployment	-0.04692	0.01898	-0.03361	0.01944
Number of Observations	zips = 69		N = 89,232	
	San Diego			
	Weighted Average Across Zip Codes		Overall County	
	Coeff. Est.	Std. Dev	Coeff. Est.	Std. Err.
σ_ϵ	0.25182	0.05760	0.27307	0.00229
σ_η	0.09400	0.01989	0.11516	0.01915
ρ	0.95105	0.07489	0.82278	0.04492
Number of Bedrooms	0.00505	0.04230	0.02742	0.00227
Lotsize	0.00000	0.00001	0.00000	0.00000
Square Footage	0.00031	0.00008	0.00042	0.00000
Ten year Treasury	0.07120	0.03496	0.10270	0.03622
County Population	0.33567	0.05242	0.44481	0.01007
Unemployment	-0.05382	0.02049	-0.07837	0.02834
Number of Observations	zips = 87		N = 71612	
	Los Angeles			
	Weighted Average Across Zip Codes		Overall County	
	Aver. Coeff. Est	Std Dev	Coeff. Est.	Std. Err.
σ_ϵ	0.24277	0.05964	0.28220	0.00146
σ_η	0.10048	0.02507	0.07939	0.01683
ρ	0.91671	0.14225	0.96506	0.03170
Number of Bedrooms	0.01125	0.02756	0.07913	0.00150
Lotsize	0.00001	0.00001	0.00000	0.00000
Square Footage	0.28303	0.07101	0.52943	0.00172
Ten year Treasury	0.08504	0.04125	0.06861	0.02816
County Population	0.35645	0.0471	0.35685	0.10174
Unemployment	-0.05884	0.02527	-0.05467	0.02030
Number of Observations	zips = 237		N = 211989	

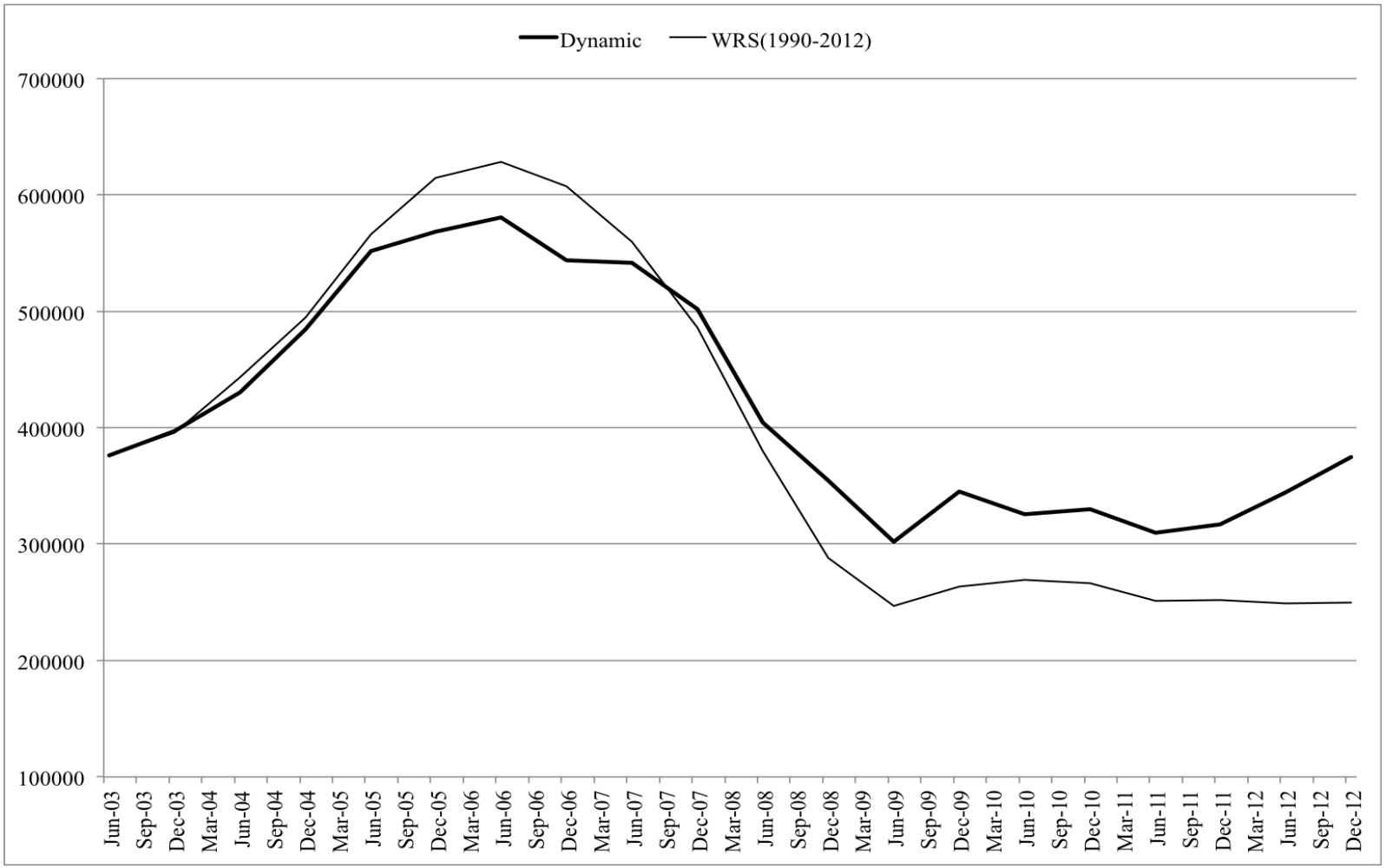
Index: Alameda



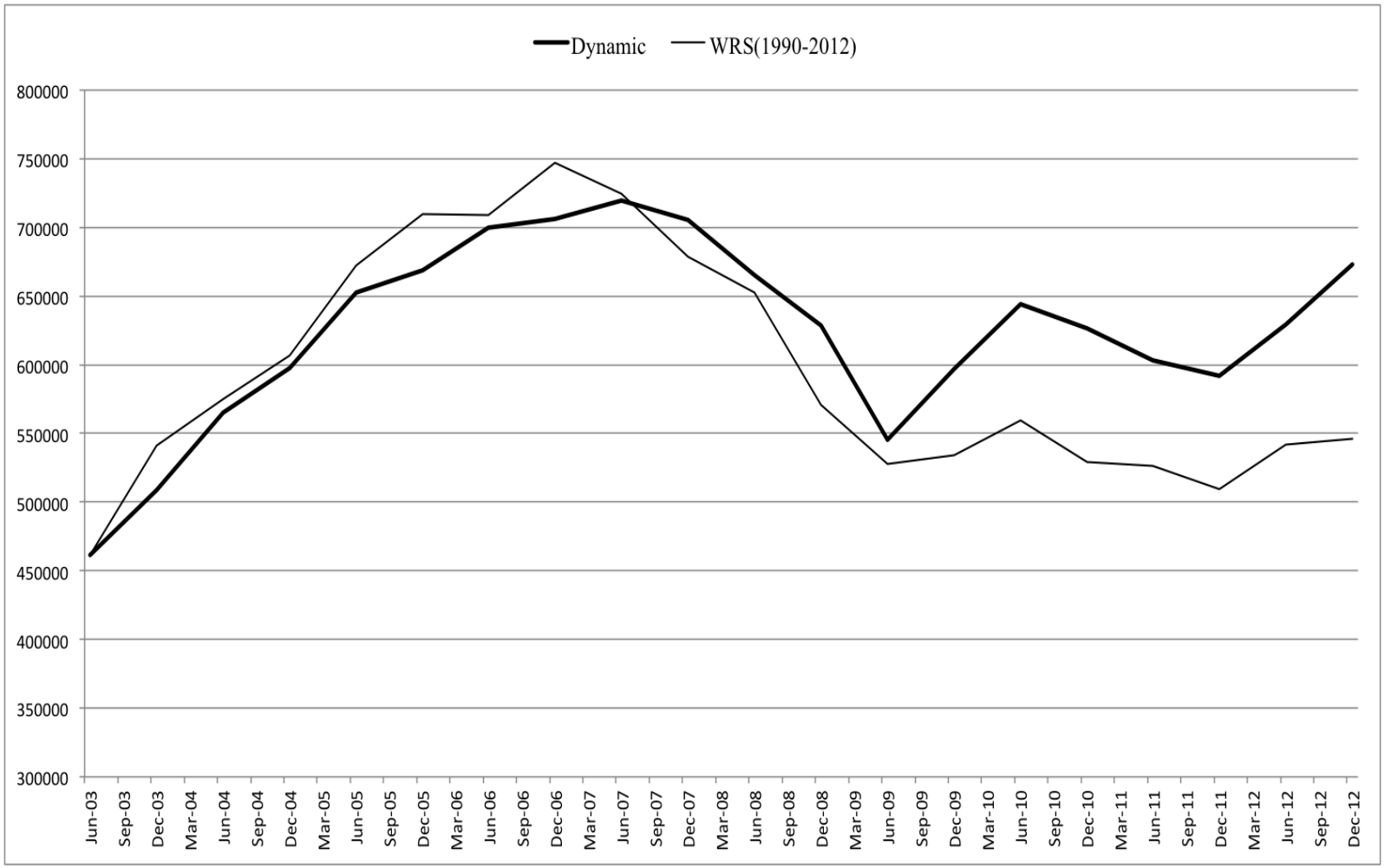
Index: San Francisco



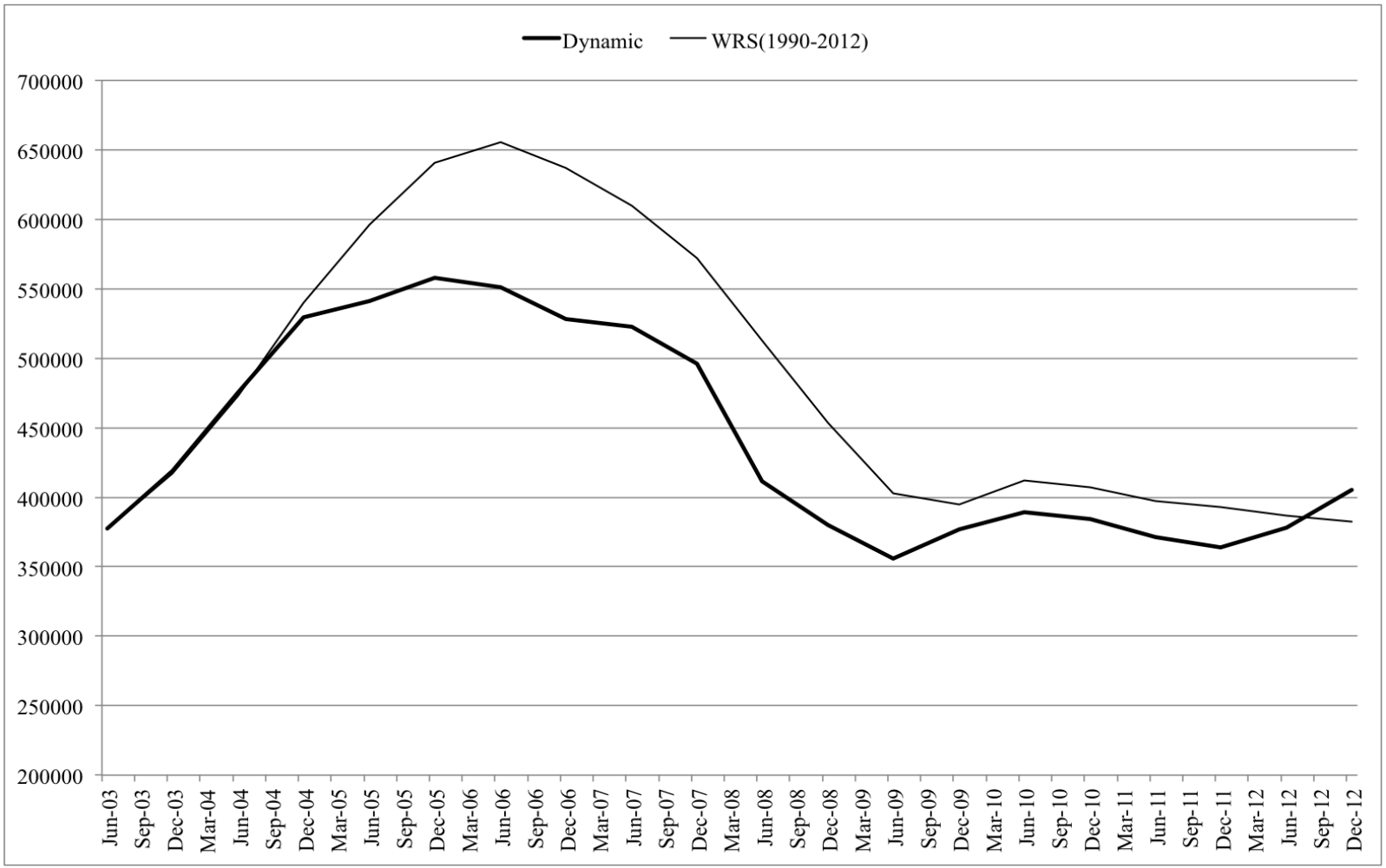
Index: Contra Costa



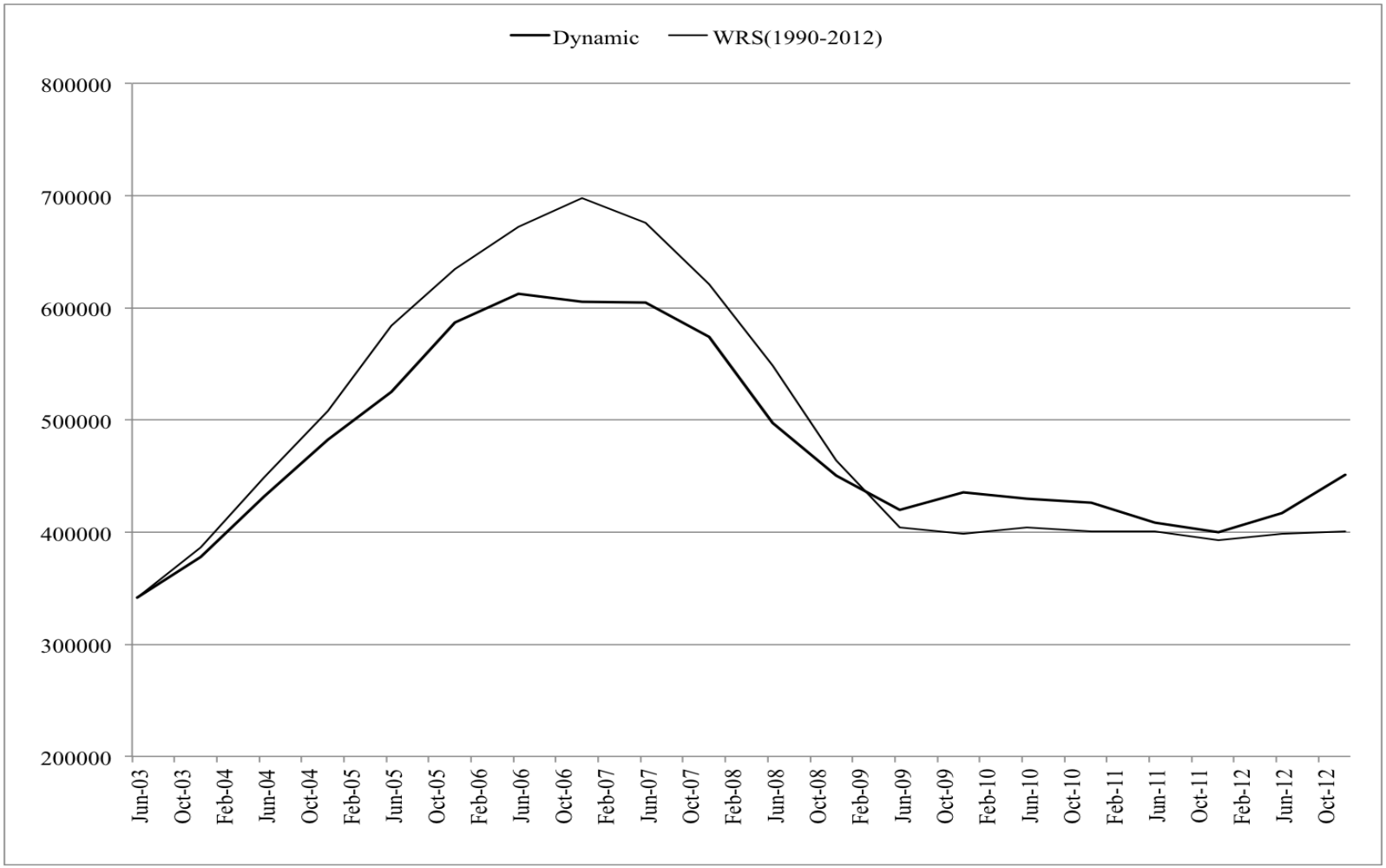
Index: Santa Clara



Index: San Diego



Index: Los Angeles



Conclusions

- Successfully estimated a new and flexible dynamic house price index suitable for mortgage valuation, stress testing, analyses of housing market price formation.
- Applied new data set, which can be extended to the whole U.S. and contains time series of property characteristics.
- Idiosyncratic house price volatilities are much larger than the often-reported volatilities of the Case-Shiller repeat-sales indices.