Aggregation bias and the repeat sales price index

Anthony Pennington-Cross[1]

Introduction

A house price index is by definition a summary indicator of spatial and/or intertemporal house prices. House price indices provide a basis for measuring real estate values and their growth through time. But, all housing is not created equal. The attributes of the home (the square feet, number of baths, quality of materials, etc) as well as the location of the home add substantial heterogeneity to the value of housing in any location. As a result, any index will measure individual house prices with an error and is best thought of representing overall market conditions. This is even true for house price index estimates at a detailed level of geography such as census tracts or zip codes.

The objective of a house price index is to accurately describe the level or change in prices for a location. In the United States, house prices are typically reported for metropolitan areas or states. For instance, the National Association of Realtors (NAR) reports median house prices for a range of metropolitan areas. In addition, the Office of Federal Housing Enterprise and Oversight (OFHEO) reports a constant quality house price index for all metropolitan areas and states. The index attempts to hold quality constant by measuring the average growth in house prices using only multiple transactions associated with the same home.

Because housing is a local phenomenon and heterogeneous in space and across time, these measures of house prices provide a highly aggregated view of house prices. As a result there is substantial evidence of heterogeneous price appreciation and sample selection issues when estimating house price indices (Dreiman and Pennington-Cross (2004), Englund et al (1998), Gatzlaff and Haurin (1997)). In addition, housing is a unique commodity because it trades infrequently. This is in contrast to other markets such as commodities, stocks, and bonds which have active centralised markets that establish market clearing prices through multiple transactions each business day. There are even intraday markets that are used to promote transactions and non-business day pricing estimates. In the housing market, if a home sells only once a year it would be extremely unusual. In fact, it would be impossible, given the time required to sell a home, for a home to sell everyday. As a result, transactions are sparse relative to the outstanding stock of homes.

Both the NAR and OFHEO price indices are best described as transaction-based house price indices. The question examined in this paper is whether transaction-based house price indices differ from true or housing stock-based house price indices.

Motivation

Consider the following, stylised representation of the housing market. This presentation focuses on the importance of differences between transactions and the stock of housing and how these differences can impact house price estimates. In a region there are two cities, *A* and *B*, with housing stock of Q_A and Q_B. The total housing stock is $Q = Q_A + Q_B$. For simplicity assume that all homes are identical within each city and that the housing stock and housing quality are time invariant. Also assume that there is no noise or a stochastic process associated with house prices. House prices in City *A* and City *B* are P_{At} and P_{Bt} in each time period t. Therefore, the prices and their growth through time within

[1] The views expressed in this research are those of the author and do not represent policies or positions of the Office of Federal Housing Enterprise Oversight or other officers, agencies, or instrumentalities of the US Government. Anthony Pennington-Cross, Office of Federal Housing Enterprise Oversight, 1700 G Street, 4th Floor, Washington, DC 20052; e-mail: apennington@ofheo.gov or anmpc@yahoo.com; tel: 202-414-8937.

each city is the same for all houses. The only difference between the two cities is how much housing stock is in each city, the price of housing in each city, and the appreciation rate of house prices through time. The region's average or true house price is defined as:

$$P_t = (Q_A/Q) \times P_{At} + (Q_B/Q) \times P_{Bt} \qquad (1)$$

Each city's price is weighted by the city share of the housing stock. The change in house prices over time can also be expressed as:

$$\Delta P_t = (Q_A/Q) \times \Delta P_{At} + (Q_B/Q) \times \Delta P_{Bt} \qquad (2)$$

Note again that each city's price is weighted by the city share of the housing stock. ΔP_t can be viewed as an index.[2] In contrast, for an index based only on observed transactions, ΔP_{Tt}, a different weighting scheme applies. A transaction-based index can be represented as:

$$\Delta P_{Tt} = (Q_{TAt}/Q_{Tt}) \times \Delta P_{At} + (Q_{TBt}/Q_{Tt}) \times \Delta P_{Bt}, \qquad (3)$$

where Q_{TAt} is the total quantity of city A's housing stock that transacted, Q_{TBt} is the total quantity of city B's housing stock that transacted, Q_{Tt} is the total amount of housing stock transacted and is defined as $Q_{TAt} + Q_{TBt}$, and ΔP_{Tt} is the transaction-based index. The transaction quantities are bounded by zero and the quantity of available housing stock. Therefore, $Q_{TAt} < Q_A$, $Q_{TBt} < Q_B$, and $Q_{Tt} < Q$. In contrast to the quantity of housing, which is held constant by assumption, the quantity of housing that transacts can also vary through time. The observed transactions, or prices, are not weighted by the share of the housing stock they represent, but instead by the share of total transactions. As a result, under certain conditions the transaction-based index can be the same or deviate from the true index.

$$(Q_A/Q) = (Q_{TAt}/Q_{Tt}) \quad \text{and} \quad (Q_B/Q) = (Q_{TBt}/Q_{Tt}) \Rightarrow \Delta P_{Tt} = \Delta P_t,$$

$$\Delta P_{At} = \Delta P_{Bt} \Rightarrow \Delta P_{Tt} = \Delta P_t \qquad (4)$$

For example, if the propensity to transact equals the fraction of the housing stock in each city then the transaction and true index will be the same. In addition, if prices increase at the same rate in both city A and city B, regardless of the propensity to transact, then the transaction and true indices will be identical.

But, when city prices increase at different rates and the propensity to transact differs then the transaction index will diverge from the actual index. Assume that homeowners are more likely to sell their homes when prices are increasing. For example, if prices are increasing faster in city A than city B and the propensity to transact is also higher in city A then the true and transaction-based indices will deviate.

$$\text{If } \Delta P_{At} > \Delta P_{Bt} \quad \text{and} \quad (Q_{TAt}/Q_{Tt}) > (Q_{TBt}/Q_{Tt}) \Rightarrow \Delta P_{Tt} > \Delta P_t \qquad (5)$$

In this scenario, using the transaction index, the price index will be estimated to be increasing at an artificially high rate. This is the source of the systematic bias in the transaction-based index. The opposite bias would be found if transactions are less likely to occur in higher appreciating locations. Supporting the first hypothesis, Genesove and Mayer (2001) found some evidence that homeowners do not like to sell their homes for a loss and are therefore less likely to transact when prices are down and more likely to transact when prices are up. This indicates that locations with robust housing markets may receive too much weight leading to a systematic upward bias in the transaction-based index. In contrast, Redfearn (2003) has found that transaction rates are sometimes positively and sometimes negatively correlated with house price movements in Sweden.

[2] As explained in the following sections, the index does not provide any information on the level of house price. Instead, for all locations, the index is normalised to one or 100 in the initial period and the growth rates derived from the resulting index.

Repeat sales models

The following section introduces a repeat sales model of house price appreciation rates to examine empirically the impacts of any systematic bias caused by using transactions to estimate average appreciation rates. This section will initially explain the repeat sales approach, which is implicitly a transaction-based index, and then introduces a new weighting scheme based on housing units to approximate the "true" or population wide price index.

Repeat sales models attempt to hold quality constant by examining only properties with repeat transactions to estimate average appreciation rates for particular locations. In this paper we include estimates at the state level. This will help to introduce a variety of appreciation rates across different cities within a single state. The house price index preserves the intuitively simple interpretation of any index. For example, if the index is 100 in state *j* in 2000 and increases to 105 in state *j* in 2001, the average house price in state *j* increased by 5% over the period 2000-01. The basic procedure dates back to Bailey et al (1963) and has remained essentially the same for over 40 years as is evidenced by Dreiman and Pennington-Cross (2004). Following the approach utilised by Case and Shiller (1987) and later modified by Abraham and Schauman (1991). It is assumed that the natural logarithm of price, P_{it}, of an individual house *i* at time *t*, can be expressed in terms of a market price index β_t and an individual house idiosyncratic deviation from the market index υ_t.

$$\ln(P_{it}) = \beta_t + \upsilon_t \tag{6}$$

The market index is expected to be correct on average so that $E(\upsilon_t) = 0$. This specification allows us to express the percentage change in price for house *i* which transacts in time periods *s* and *t* as:

$$\Delta V_i = \ln(P_{it}) - \ln(P_{is}) = \beta_t - \beta_s + \upsilon_t - \upsilon_s \tag{7}$$

Using D_{i_τ} a dummy variable that equals one if the price of house *i* was observed for a second time at time τ, –1 if the price of house *i* was observed for the first time at time τ, and zero otherwise the growth in house prices can be estimated by:[3]

$$\Delta V_i = \sum \beta_\tau D_{i_\tau} + \varepsilon_i, \quad \text{where } \varepsilon_i = \upsilon_t - \upsilon_s \tag{8}$$

Assuming $E(\varepsilon_i) = E(\upsilon_t) - E(\upsilon_s) = 0$, the parameters β_τ, $\tau = 0, 1, 2, \ldots, T$ for the market index can be estimated by ordinary least squares (OLS) regression.[4] Abraham and Schauman (1991) introduced the concept that the variance of the house prices around this estimated mean appreciation rate is likely to increase the longer it is between transactions. Therefore, OLS is not an efficient estimator because we cannot assume that the variance of the error term is constant. The squared deviations of observed house prices from the market index are given by:

$$\varepsilon_i^2 = (\Delta V_i - \sum \beta_\tau D_{i_\tau})^2 \tag{9}$$

It is assumed that the squared deviations of observed house price changes around β_τ will provide us with an estimate for the variance of the error term. The estimated variance of the error term will change for each combination of *s* and *t*.

$$E[\varepsilon_i^2] = A(t-s)_i + B(t-s)_i^2 + C \tag{10}$$

The expected values, from the estimate parameters *A*, *B*, and *C* and $t-s$, of the squared deviations, $E[\varepsilon_i^2]$, are used to derive the expected standard error, $E(se_i)$, which is defined as the square root of $E[\varepsilon_i^2]$. The expected errors are then used as the weights needed to obtain GLS estimates of the B_τ parameters in the following regression:

$$\Delta V_i / E[\varepsilon_i^2] = \sum \beta_\tau D_{i_\tau} / E[\varepsilon_i^2] + \varepsilon_i / E[\varepsilon_i^2] \tag{11}$$

[3] Note that the time period τ, which indicates the time period for which the index is estimated, is different from *t*, which was used previously to denote the time period of the second transaction.

[4] It is necessary to restrict one of the market index parameters to avoid perfect co-linearity among the explanatory variables. It is convenient to use $\beta_r = 0$, where *r* is the base period of the reported index.

This specification is estimated to derive house price indices. Index numbers for periods $\tau = 1, 2, 3, \ldots, T$ are given by:

$$I_\tau = 100e^{\beta_\tau^*} \qquad (12)$$

where β_τ^* are the GLS parameter estimates of the market index.[5] The market index is a transaction based index because it only includes properties that transacted. If there are 1,000 observed repeat transactions then there are 1,000 observations in the estimation data set. Each observation is implicitly weighted equally. As hypothesised in the previous section, the propensity for a house to transact may be positively correlated with increasing house prices. If this is true, then transactions in locations with rising house prices represent less housing stock than transactions in locations where house prices are not increasing as much or declining. Therefore, the implicit equal weighting used to estimate the transaction-based market index is inaccurate and would bias the estimates from the true appreciation rate.

To create a housing-stock based or true market index, each observed change in house price (from the repeated observations) is weighted by the fraction of the housing stock in the neighbourhood. In this paper, the index estimated is at the state level and census tracts define the neighbourhoods. The US Census Bureau reports housing units in each tract in census years from www.census.gov, for download by county. The weights are defined using the 1990 and 2000 census tract housing units data. Because the transactions can span a considerable time period a decision rule is developed to assign the correct weight: (1) If both transaction are prior to 1991 then the 1990 census weights are used, (2) If both transactions are after 2000 the 2000 census weights are used, (3) If one of the transactions occurred during the years 1991 through 1999, then the median year of the period in which the loan was alive is used. The median year is used to identify the weight to be used from a straight-line spline of the 1990 and 2000 weights.

Results

Table 1 provides a graphical representation of the estimated annual appreciation rate for house prices for six representative states (California, Massachusetts, Maryland, Missouri, Nevada, and Ohio). The six states include locations where house prices have experienced large cycles (California and Massachusetts), locations where prices have been fairly stable through time (Ohio and Missouri), and a smaller state with a dominant and growing metropolitan area (Nevada). Some states such as Nevada or Missouri are dominated by one or two cities. In contrast, California includes a wide variety of cities with vastly different types of economies ranging from agricultural economies to high tech and financial economies. This heterogeneity should help to create deviations in house price appreciation rates and deviations in the propensity to transact. These are the conditions identified as ingredients that should make the transaction-based index deviate from the true index.

In contrast to the theory, the results provide very little evidence of any aggregation bias associated with the transaction based sample. For instance, in California there is almost no discernable difference between the index using transaction weights and the one using housing stock weights. Recall that one plausible hypothesis was that the propensity to transact should increase the more house prices are rising in a particular location. This should help to create a divergence of the transaction-based index and the housing stock based index if the propensity to transact is procyclical. But, in California there is almost no difference between the two indices, proving little support for the theory.

The same is true in Massachusetts, another location that has experienced a large run-up in house prices during the mid-1980s, price deflation and stagnation from 1988 through 1993 and modest inflation until the end of the time period. Again in this scenario, assuming heterogeneity in transaction propensities the indices should diverge. Instead, the transaction and housing stock indices are almost identical.

[5] If the restriction $\beta_1 = 0$ is imposed in estimation, then $I_1 = 100$.

The state of Maryland is substantially smaller, but is dominated by Washington DC suburban neighbourhoods and Baltimore. Again, there is almost no difference between the transaction and the housing stock based indices.

Ohio also experienced the run-up in house prices from 1985 through 1987, but the magnitude of the increases was much smaller than for Maryland, Massachusetts or California. In, contrast though, Ohio has not experienced any declining prices, but has roughly held at a 3% appreciation rate from 1990 through the end of 2000. Despite these different housing market experience the two indices are, again, almost identical.

In the two remaining states (Nevada and Missouri) the transaction and housing stock indices do diverge. In both states the peak of the run-up in house prices is over-stated in the transaction index. This is apparent in Nevada during in 1988 and 1989 and in Missouri 1986 as well as in 2000 for both states. Nevada is a unique state because the rapid growth of Las Vegas throughout the 1990s and the relative abundance of developable land in the desert. In contrast, Missouri's housing market is dominated by St Louis, which is a city that has experienced a steady decline in population. But the area still includes some major employers such as a several large mortgage corporations. The deviations are much larger in Nevada and are especially apparent from 1992 through 1994 when house price growth was moderating after larger increases in the late 1980s. In fact, the housing stock index smoothes the transaction index. The results in Nevada are not consistent with a procyclical propensity to transact theory. Instead they indicate that in Nevada the propensity to transact was higher in locations with faster increasing prices during the price run-up in the late 1980s. But during the price decline/stagnation of the early 1990s the propensity to transact was higher in neighbourhoods experiencing the worst declines in prices.

In summary, there is no consistent evidence supporting the need for focus on housing stock rather than transactions when creating a repeat sales house price index or the existence of a procyclical propensity to transact across cities.

Home owner negative equity

For an individual home, i, the probability of negative equity, π, can be calculated as follows:

$$\pi_{\tau,t-s} = \Theta((\log \text{upb}_{t-s} - \log P_t)/(E(se_{t-s})) \qquad (12)$$

where $\pi_{\tau,t-s}$ is the probability that the property is worth less than the mortgage and depends on the τ, the current time period, as well as how long it has been since the last transaction ($t-s$), upb_{t-s} is the unpaid balance on the mortgage and depends on how long the borrower has been paying the mortgage, P_τ is the value or price of the home, $E(se_i)$ is the expected or estimated standard error from equation (10), and Θ is the cumulative normal density function (see Pennington-Cross (2004), Deng (1997), Deng et al (1994)).[6] Assume that the mortgage interest rate is fixed at 8% for the life of the loan, the term is fixed at 30 years, the home initial value is 100 dollars, and a 10 dollar down payment was made. In addition, the borrower is assumed to make all payments on time so that the unpaid balance is reduced on schedule through the 30 years. Lastly, to isolate the impact of new price index estimate from the impact of the standard error estimates assume that prices in all states are constant at 100.

Using these assumptions Figure 2 shows the difference between the transactions estimated π and the housing stock based π. For instance, if the transaction $\pi = 7\%$ and the housing stock $\pi = 8\%$ the percent deviation is 1%. For all states, except Nevada, the deviations reported for the first five years of the mortgages life is always negative and always less than 1%. In Nevada the deviations are positive and can exceed 3%. Therefore, while the dispersion of house prices around the mean is usually larger using the transaction index, the dispersion estimates are very similar in terms of overall magnitude. This leads to a slight overestimate of the probability that the borrower has negative equity. Again, in Nevada the results are the opposite.

[6] The expected variance is time varying as defined by the parameter estimates of A, B, C and the time between transactions ($t-s$).

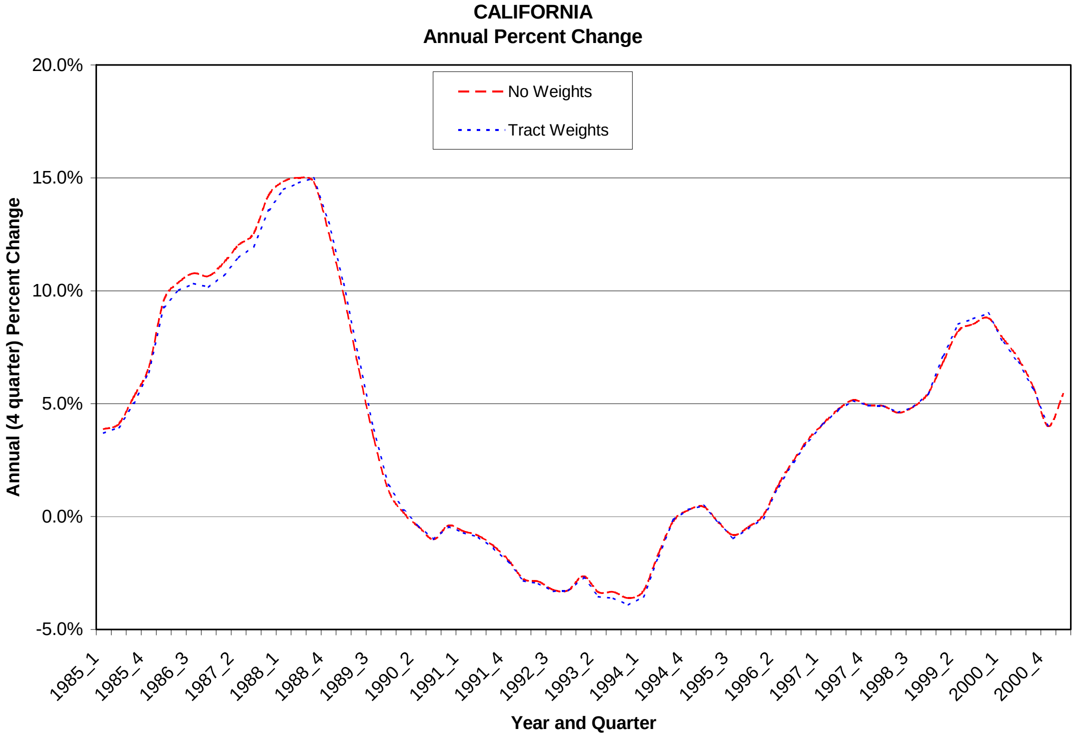
Conclusion

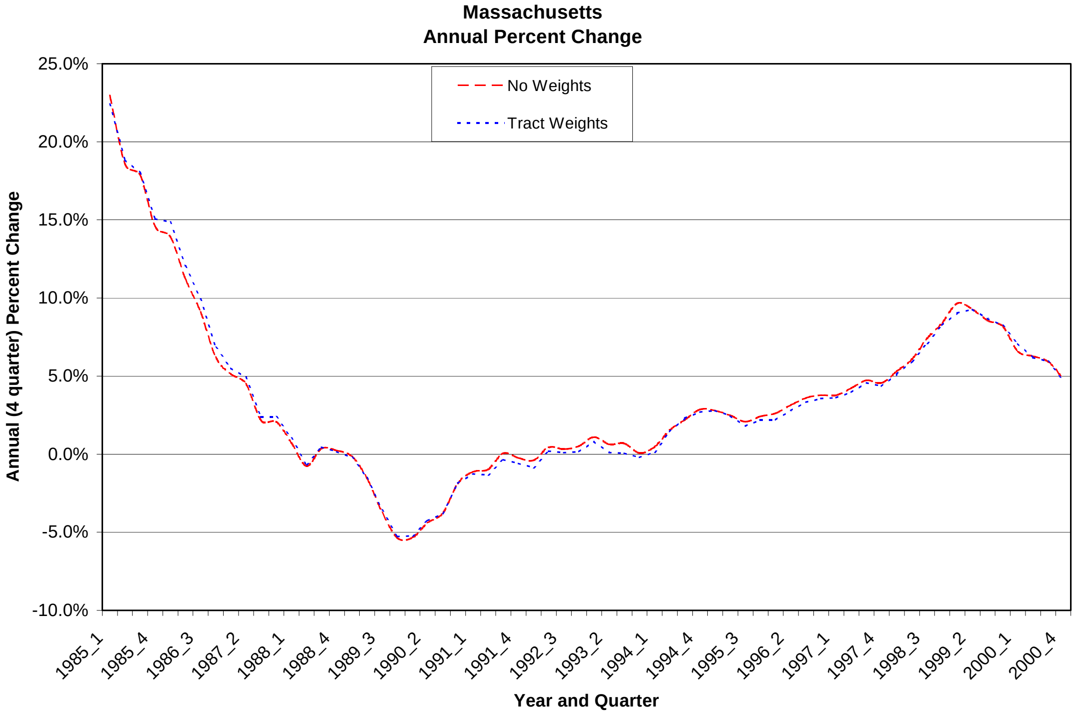
The construction of any price index must rely on actual transactions to create the index. By construction the index is an aggregate representation of individual prices. This aggregation contains a variety of property types and neighbourhood types. It is unlikely that all neighbourhoods experience the same appreciation rates or the same propensity to transact. As a result of this heterogeneity the construction of a transaction-based index may suffer from asymmetric appreciation and selection issues, which could bias the house price index.

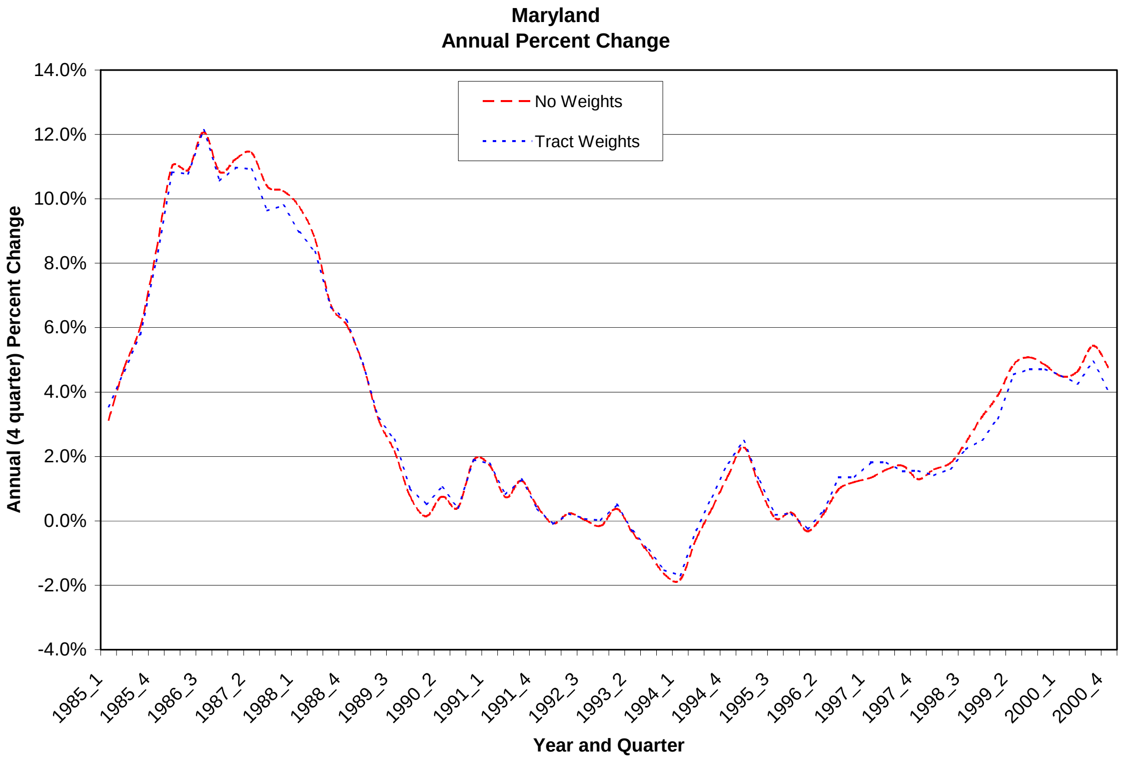
This paper examines whether any consistent bias can be found in the creation of a repeat sales price index at the state level. This is done by comparing a transaction-based index with a housing-stock-based index. The housing-stock-based index weights each observed repeat transaction by the amount of housing it represents. Therefore, the aggregate or regional index should reflect the true appreciation of house prices. But, the empirical results do not indicate any substantial revisions in the index nor do the results show any large differences on the dispersion of individual house prices around the mean appreciation rate. In particular, in large states and in states that have experienced strong housing cycles almost no discernable difference between the two indices is apparent.

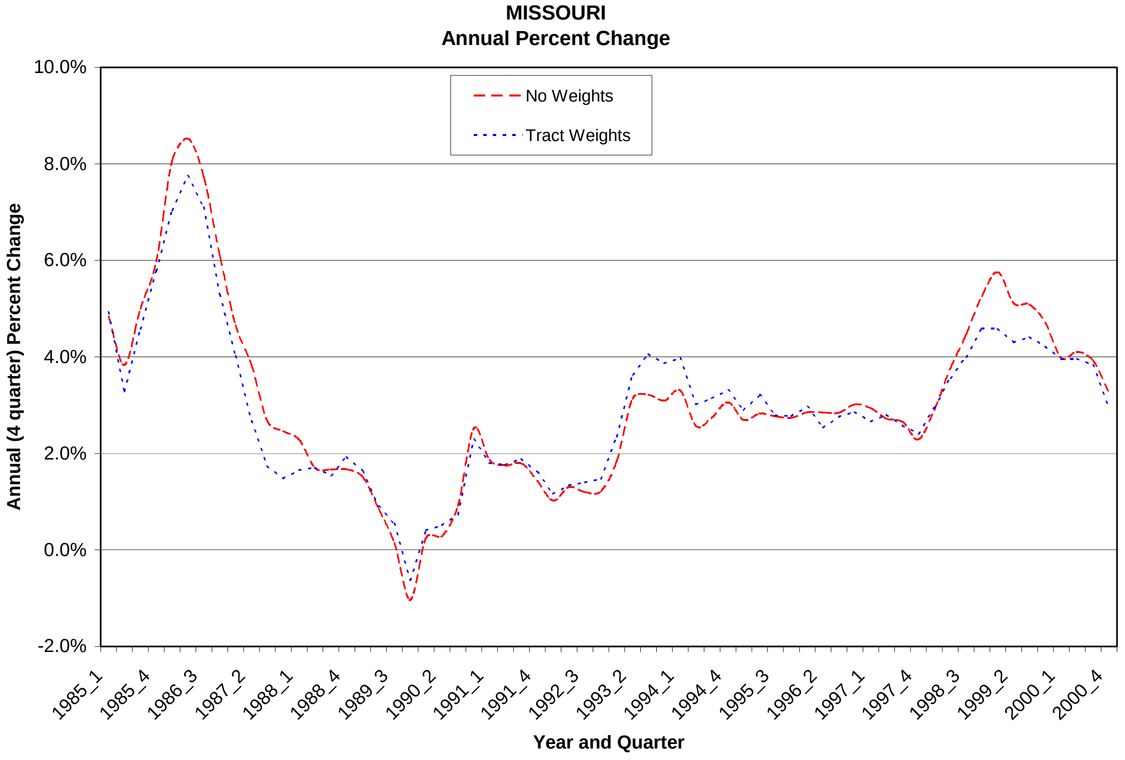
Figure 1

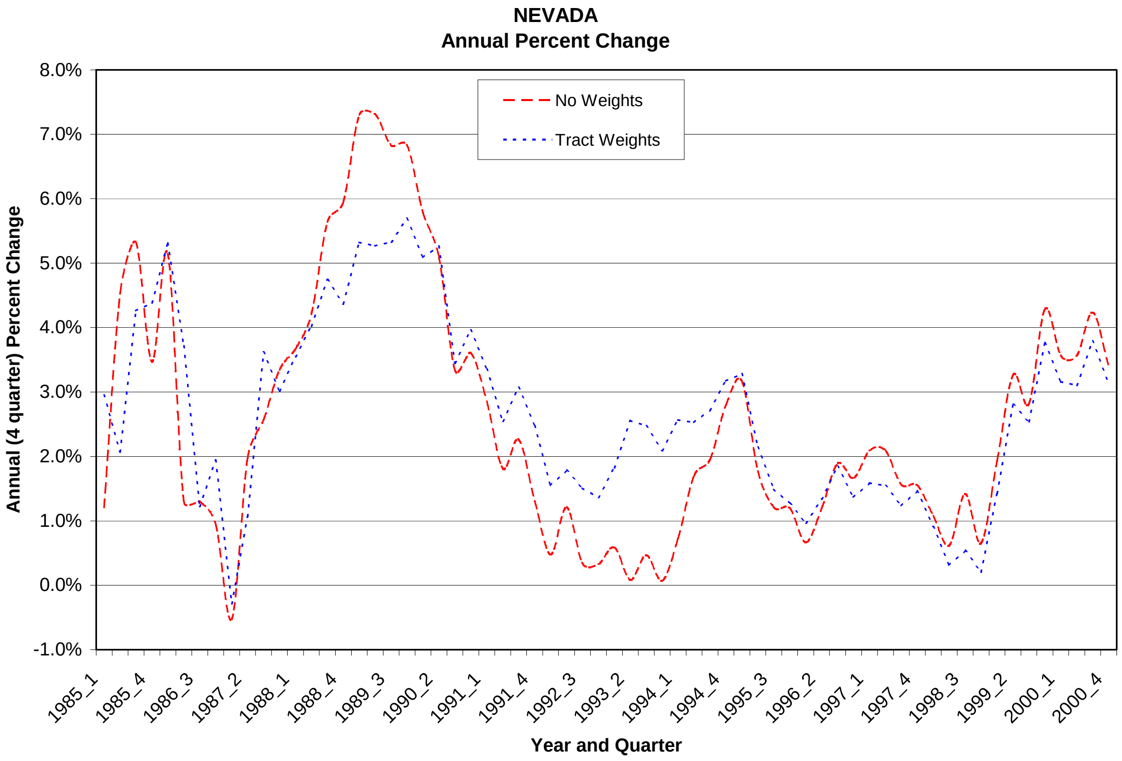
Index comparisons






Maryland
Annual Percent Change
No Weights
Tract Weights
14.0%
12.0%
10.0%
8.0%
6.0%
4.0%
2.0%
0.0%
-2.0%
-4.0%
Annual (4 quarter) Percent Change
1985_1
1985_4
1986_3
1987_2
1988_1
1988_4
1989_3
1990_2
1991_1
1991_4
1992_3
1993_2
1994_1
1994_4
1995_3
1996_2
1997_1
1997_4
1998_3
1999_2
2000_1
2000_4
Year and Quarter


MISSOURI
Annual Percent Change
No Weights
Tract Weights
10.0%
8.0%
6.0%
4.0%
2.0%
0.0%
-2.0%
Annual (4 quarter) Percent Change
1985_1
1985_4
1986_3
1987_2
1988_1
1988_4
1989_3
1990_2
1991_1
1991_4
1992_3
1993_2
1994_1
1994_4
1995_3
1996_2
1997_1
1997_4
1998_3
1999_2
2000_1
2000_4
Year and Quarter


NEVADA
Annual Percent Change
No Weights
Tract Weights
8.0%
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
-1.0%
Annual (4 quarter) Percent Change
1985_1
1985_4
1986_3
1987_2
1988_1
1988_4
1989_3
1990_2
1991_1
1991_4
1992_3
1993_2
1994_1
1994_4
1995_3
1996_2
1997_1
1997_4
1998_3
1999_2
2000_1
2000_4
Year and Quarter

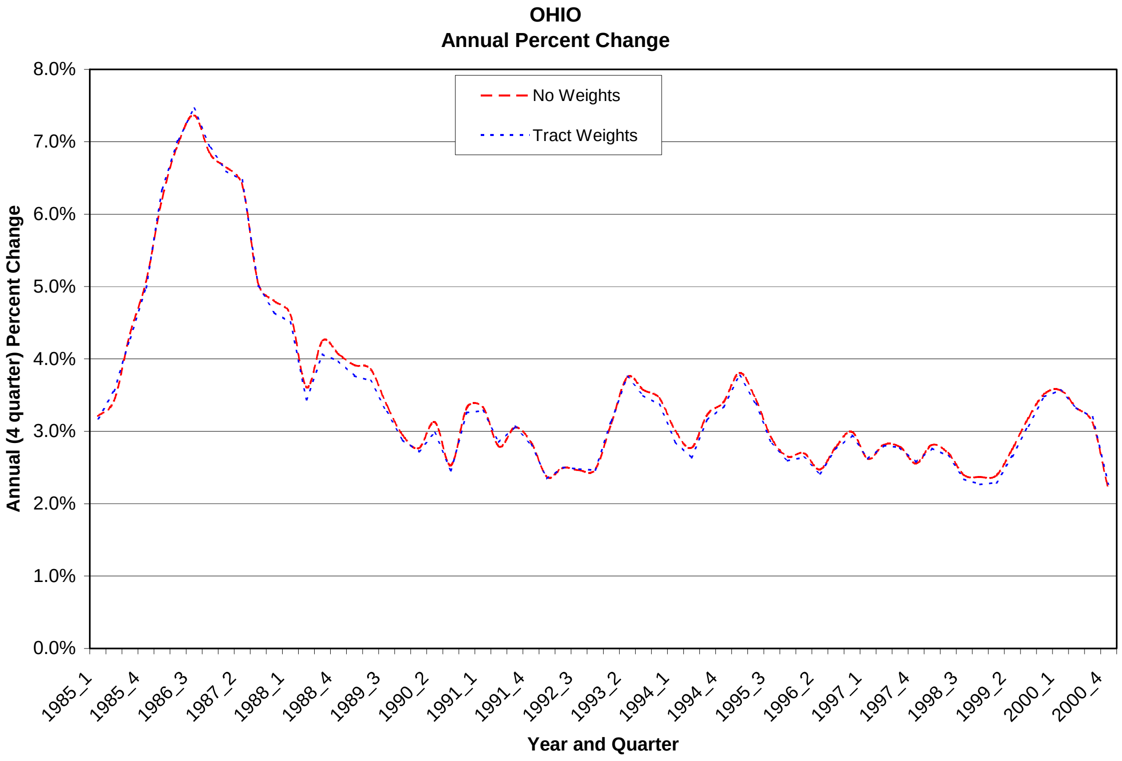
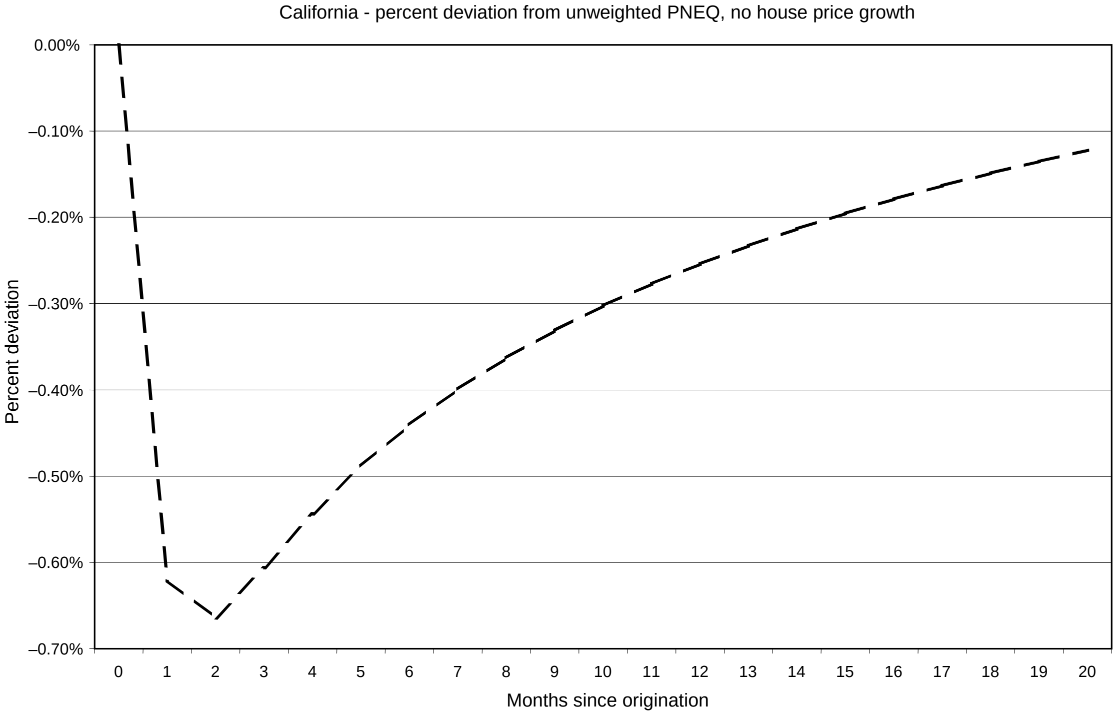
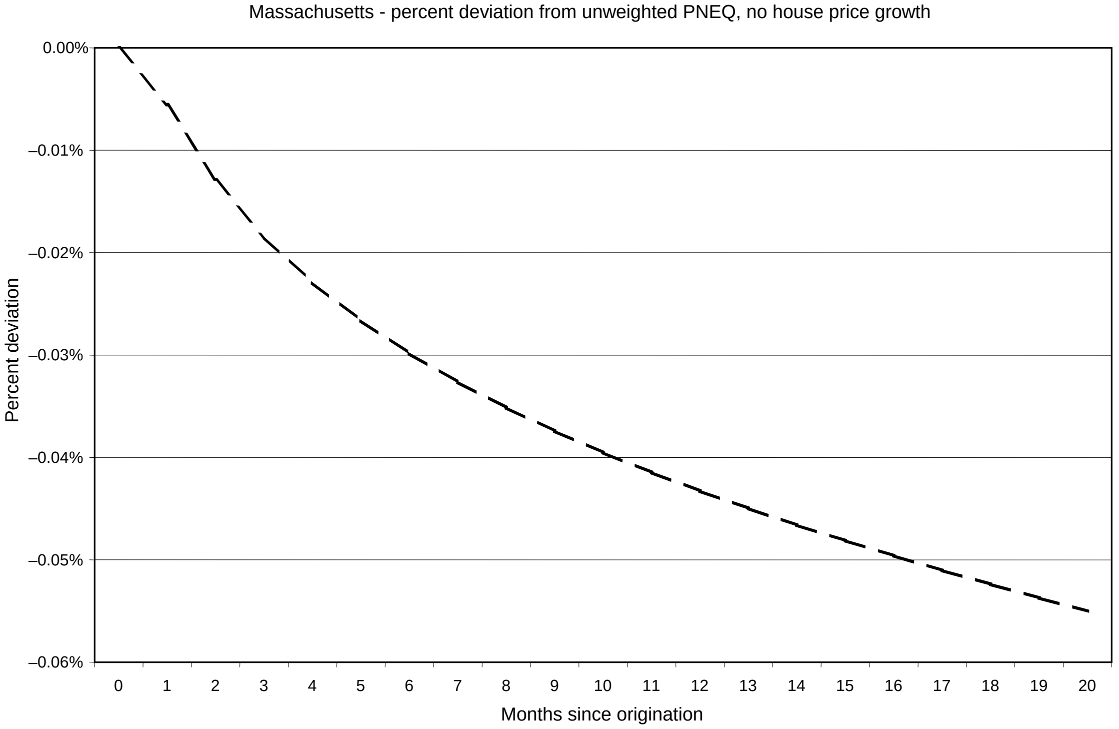

OHIO
Annual Percent Change
No Weights
Tract Weights
8.0%
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
Annual (4 quarter) Percent Change
1985_1
1985_4
1986_3
1987_2
1988_1
1988_4
1989_3
1990_2
1991_1
1991_4
1992_3
1993_2
1994_1
1994_4
1995_3
1996_2
1997_1
1997_4
1998_3
1999_2
2000_1
2000_4
Year and Quarter

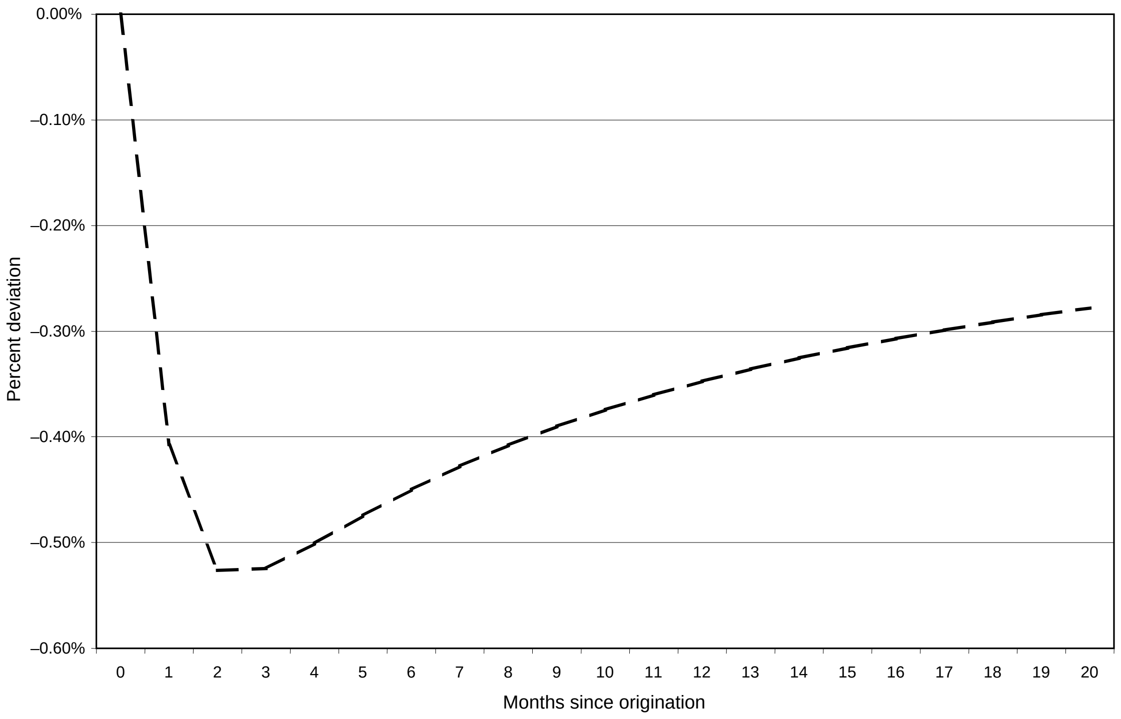
Figure 2

PNEQ deviations

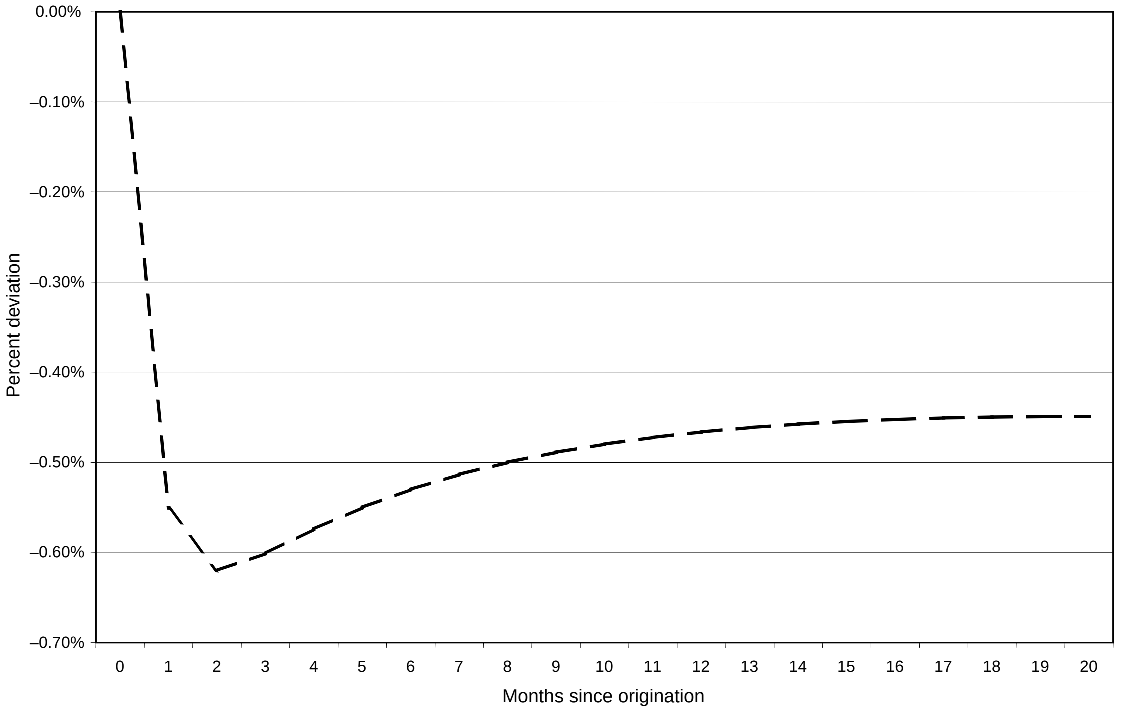




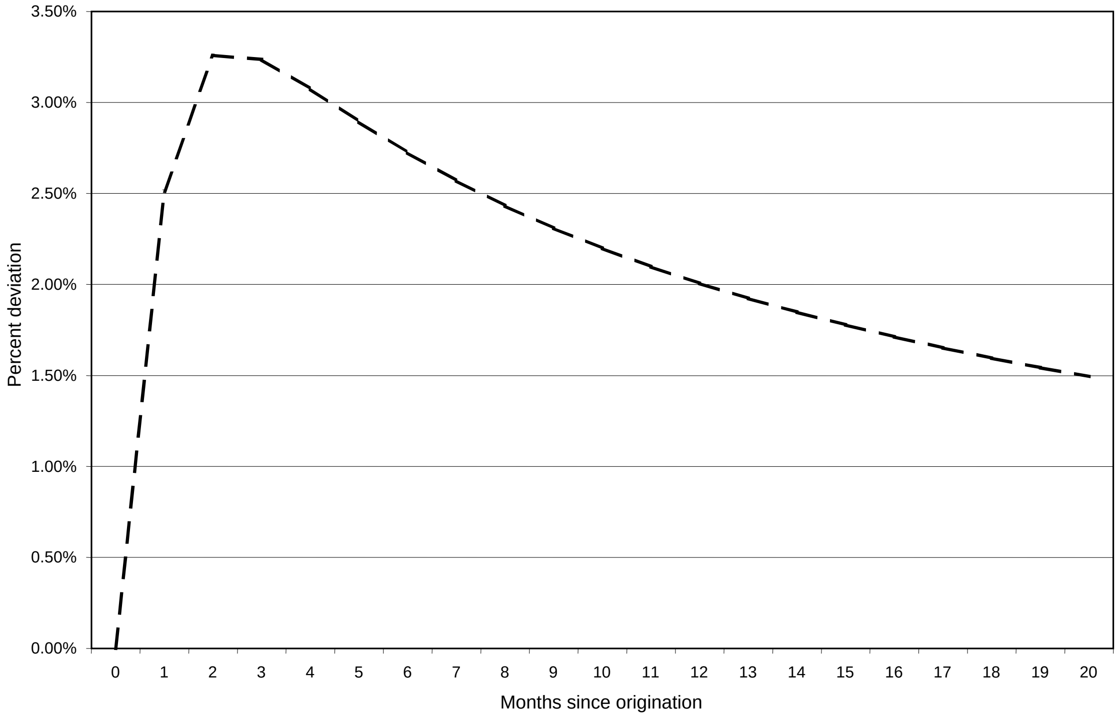
Maryland - percent deviation from unweighted PNEQ, no house price growth



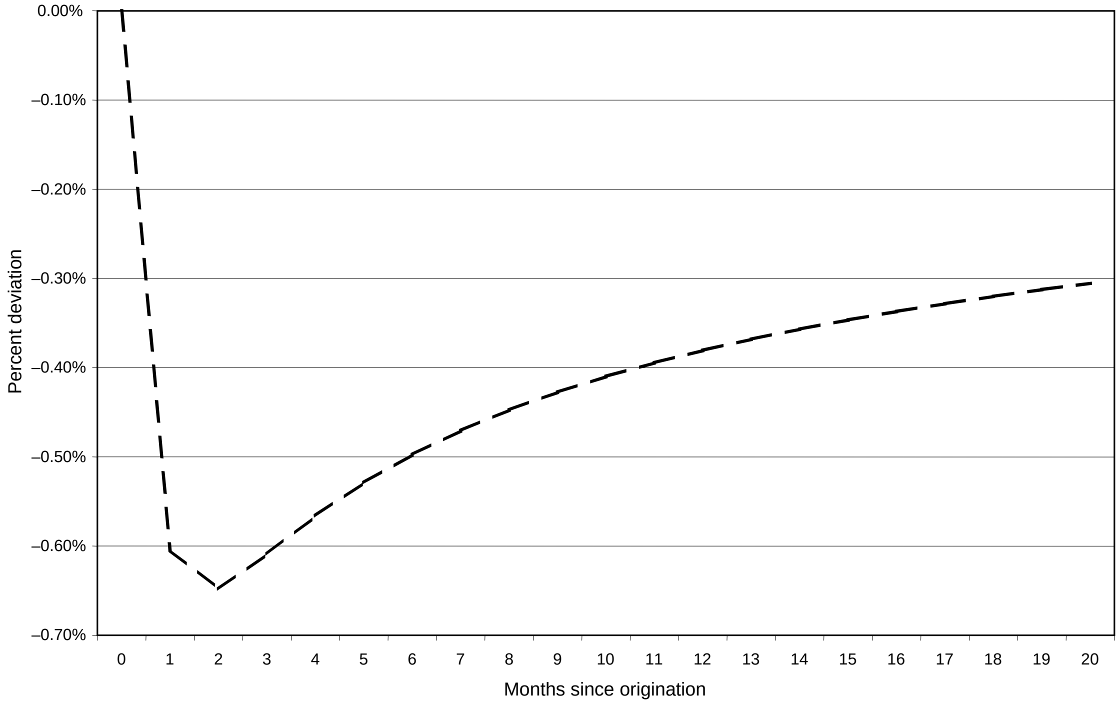
Missouri - percent deviation from unweighted PNEQ, no house price growth



Nevada - percent deviation from unweighted PNEQ, no house price growth



Ohio - percent deviation from unweighted PNEQ, no house price growth



References

Abraham, Jesse and William Schauman (1991): "New evidence on home prices from Freddie Mac repeat sales", *AREUEA Journal*, 19(3), pp 333-52.

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, pp 933-42.

Case, Karl and Robert Shiller (1987): "Prices of single family real estate", *New England Economic Review*, pp 45-56.

Deng, Youngheng (1997): "Mortgage termination: an empirical hazard model with stochastic term structure", *Journal of Real Estate Finance and Economics*, 14(3), pp 309-29.

Deng, Youngheng, John Quigley and Robert Van Order (1994): "Household income, equity, and mortgage default risks", working paper, University of California-Berkeley.

Dreiman, Michelle and Anthony Pennington-Cross (2004): "Alternative methods of increasing the precision of weighted repeat sales house prices indices", *Journal of Real Estate Finance and Economics*, forthcoming.

Englund, Peter, John Quigley and Christian Redfearn (1998): "Improved price indexes for real estate: measuring the course of swedish housing prices", *Journal of Urban Economics*, 44(2), pp 171-96.

Gatzlaff, Dean and Donald Haurin (1997): "Sample selection bias and repeat-sales index estimates", *Journal of Real Estate Finance and Economics*, 14(1), pp 33-50.

Genesove, David and Christopher Mayer (2001): "Loss aversion and seller behavior: evidence from the housing market", *The Quarterly Journal of Economics*, 116(4), pp 1233-60.

Pennington-Cross, Anthony (2004): "Credit history and the performance of prime and nonprime mortgages", *Journal of Real Estate Finance and Economics*, 27(3).

Redfearn, Christian L (2003): *Think globally, aggregate locally: index consistency in the presence of asymmetric appreciation*, presented at the American Real Estate and Urban Economics Association January sessions.