House Price Tiers in Repeat Sales Estimation

Douglas McManus[1]
Freddie Mac, 8200 Jones Branch Drive, McLean, VA 22102

Abstract

The existing methods for estimating tiered house price indexes are subject to substantial biases because a property's tier can only be measured imprecisely. Both academic research and industry practice have implemented tiered house price index estimation techniques but without a methodological solution to the bias problem. This paper proposes bootstrap procedures for correcting this bias in the context of testing for the existence of house price tiers. This method is illustrated at the state level for California, and it is shown that there are statistically significant tier effects.

Key Words: House price modeling, bootstrap, tiered markets

1. Introduction

There is a need for an econometrically valid approach to for testing for and estimating tiered house price paths. Frequently residential real estate markets are segmented into distinct sub-markets by property value. House price trends in the market for lower priced ‘starter homes,’ that are small and lack many amenities, will be influenced by economic factors that affect first time home buyers, such as credit availability. Mid-market price trends may be influenced by recent house price appreciation as many homeowners looking to ‘trade-up’ rely on equity in their existing homes as a down payment. High valued properties’ demand may be influenced by factors such as executive bonuses and stock market valuations. Economic factors may influence supply and demand differentially across these value tiers, resulting in distinct house price tier trajectories.[2]

A number of approaches to estimating tiered house price indexes exist. In each case, a rule for classifying observations into tiers is applied and then the repeat sales model is fit separately to each subset of data. Leventis (2012) identifies four different classification rules used to assign tiers to repeat transactions and shows that all of the approaches lead to house price indexes with significant bias. The bias in the index estimators will render statistically invalid any investigation of house price tiers using these methods.

The main contribution of this paper is creating a bootstrap bias correction for the bias introduced through tiering. The bootstrap has been shown to be an important and effective technique in addressing a wide range of econometric problems. The core idea of the bootstrap approach is to approximate the true but unknown data generation process by the sample, and then evaluate the statistical properties of an estimator or test statistic through resampling techniques applied to the sample.[3] In this paper, the bootstrap is utilized to estimate the bias in the tiered index and subtracting this estimated bias from

[1] Any views expressed are solely those of the author and do not necessarily represent the opinions of Freddie Mac, or its Board of Directors.

[2] Motivation for the existence of house price tiers is given in Goodman (1978), Pollakowski, Stegman, and Rohe (1991), Delaney et al. (1992), Mayer (1993), and Gatzlaff and Haurin (1997).

[3] For a discussion of bootstrap methods see Efron (1979), Hall (1992) and Horowitz (2001).

the raw tiered estimates then creates a bias-corrected index. This bias correction is utilized in the context of the statistical testing of the existence of tiers.

The specific tiering method we investigate was introduced in Porteba (1984, 1991) and used by Mayer (1993) to estimate tiered condominium prices in the Boston area. This method classifies each transaction pair based on the average of the two transaction values, where value is expressed in real terms, adjusting for house price inflation using a local house price index. Because this method blends both first and second transactions values in assigning a tier, it is expected that the bias will be partially mitigated because the biases introduced by tiers based on either the two transactions are expected to have opposite signs and hence roughly offset.[4]

This bootstrap bias correction to the repeat sales index estimator is applied to data for California. It is shown that tiering induced bias in the index is approximately linear in time; after bias correcting, the test of distinct tiers is significant; and bias–correcting increases house price growth in the low-value tier and reduces growth in the high-value tier. The middle-tier growth estimates exhibit only a very small bias.

2. The Repeat Sales House Price Model with Tiers

2.1 The Repeat Sales Model

The repeat sales model estimates a house price index based on changes in values of homes in a geographic area that have sold at least twice, and this model is typically fit using least squares.[5] The estimation involves two stages. The first-stage regression estimates the house price index using paired transaction values, and provides an index most consistent with the observed property value changes. The squared residuals from this least squares regression capture the house-specific volatility of the property value relative to the market.

In the standard repeat sales framework, house prices are presumed to follow the process:

$$ln(P_{it}) = I(t) + H_{it} + N_{it}, \qquad (1)$$

where P_{it} is the price of an individual house i at time t, $I(t)$ is a market price index that is a function of time, H_{it} is taken to be the house-specific shock to property value around the index, and N_{it} represents the house-specific noise associated with the sale of a property at a given date. The random shock N_{it} can be thought of as a house-specific pricing error, and H_{it} as driven by fundamental shocks to housing supply and demand at the property level, such as local employment losses or changes in amenities. The variance of H_{it} and

[4] The other methods discussed by Leventis (2012) are: (1) classifying each pair based on the first of the two transaction values, which is expected to bias upwards estimated growth for the low price tier properties (and downward bias the growth for the upper tier); (2) classifying each pair based on the second transaction value which is expected to induce a downward bias in estimated growth for the low price tier properties (and an upwards bias in growth for the upper tier); (3) a property based classification rule which classifies each house based on the average of the all available transaction values for a property, where each transaction value is expressed in real terms, adjusted for house price inflation using a local house price index.

[5] The original specification of this model was due to Baily, Muth, and Nourse (1963) and Case and Shiller (1987, 1989). See Wang and Zorn (1997) for a review of this model and its estimation.

N_{it} are given by σ^2_N and σ^2_H respectively. The systematic part of house price variation is captured by the variability over time in the repeat sales index *I(t)*.

As is standard in this literature, the repeat sales index, *I(t)*, is estimated through regression as a step-function with monthly increments. The dependent variable is the natural logarithm of the house price appreciation for a pair of transactions at dates *s* and *t* (assuming $s<t$), and the independent variables are a series of indicator variables representing different months that have a value of -1 and 1 at first and second transaction months respectively, and 0 otherwise. In this specification the model is estimated as:

$$log(P_t/P_s) = I(t) - I(s) + \eta, \tag{2}$$

An important driver of the bias in tiered estimation is the volatility of the error term, η, in equation (2). As equation (1) specifies, η can be written as the sum of the following components:

$$\eta = N_s + N_t + \Sigma^t_{j=s+1} H_j, \tag{3}$$

and in the absence of house price tiering and homogeneity of the variances across time, the error specification is that,

$$E(\eta) = 0 \text{ and,} \tag{4}$$

$$Var(\eta) = \sigma^2_N + \sigma^2_N + \Sigma^t_{j=s+1} \sigma^2_H. \tag{5}$$

In this context, a second-stage regression performed on the residual terms from the first-stage regression provides consistent estimates of the variance of the error term as a function of holding period. Specifically, the dependent variable is the squared residuals from by the first-stage regression and these are regressed on an intercept and the time between transactions, i.e. the holding period. The variance of the H_{it} and N_{it} shocks, i.e. σ^2_H and σ^2_N, can be inferred as from the regression coefficients,

$$Var(\eta) = 2\sigma^2_N + (t\text{-}s)\sigma^2_H \tag{6}$$

That is, the intercept is equal to twice the house-specific variance and the random walk component *(i.e.,* $\Sigma^t_{j=s+1} H_j$*)* has a variance equal to a linear function of the time between the pair of transactions. As is common in this literature, we will also include a quadratic term in holding period to better approximate the empirical form of the variance function.

2.2 Defining House Price Tiers

House price tiers are determined by classifying the paired transaction data into a discrete set of price tiers. The specific tiering method investigated classifies each transaction pair based on the average of the two transaction values, where value is expressed in real terms, adjusting for house price inflation using a local house price index. For simplicity, this paper will focus on the case of three tiers (high, medium, and low value tiers) and will denote the respective indexes by superscripts, (I^h, I^m, I^l).

Each observation (a pair of prices) is assigned to a tier by first adjusting each price for house price inflation to a specific time period, t*, using a house price index (in this implementation, estimated assuming no tiering), and then averaging the two values. Specifically, given an initial index I*, the two log prices $log(P_t)$and $log(P_s)$ are adjusted for house price inflation to date t* and averaged resulting in a value p_a:

$$p_a = (log(P_t)[I(t^*) – I(t)] + log(P_s)[I(t^*) – I(s)])/2 \qquad (7)$$

This average for each observation is compared to the distribution of all the averages in the sample to determine the tier assignment. In this paper, the sample is partitioned into three equally sized sub-samples (h, m, l) by each pair's average p_a. Denote the thresholds of p_a for the high and low tiers by p_h and p_l, which correspond to the second and first terciles of distribution. For example, if a given observation has $p_a > p_h$, then that observation will be classified as in the high tier.

A central problem with the econometric estimation of tiered estimates is that in the tiered regression equation,

$$E[log(P_t/P_s)] = I(t) – I(s) + E(\eta \mid tier = j) \text{ for tier } j, \qquad (8)$$

it will generally be the case that, $E(\eta \mid tier = j$) will no longer be zero and so the regression estimates will be biased. Further insight into this bias can be gained by analytically expressing the expectation, $E(\eta \mid tier = j)$ assuming normality of the error term. For example, for an observation to be classified as high tier it must be the case that, $p_a > p_h$, which implies,

$$(log(P_t)+ [I(t^*) – I(t)] + log(P_s)+ [I(t^*) – I(s)])/2 > p_h. \qquad (9)$$

Since $log(P_t) = log(P_s)+ [I(t) – I(s)] + \eta$, then the condition (9) can be written as $\eta > 2(p_h – (log(P_s)[I(t^*) – I(s)])$.

And so under normality, the expectation of a given observation’s error term is,

$$\begin{aligned} E(\eta \mid tier = h) &= E(\eta \mid \eta > 2(p_h – (log(P_s)[I(t^*) – I(s)])) \\ &= E(\eta \mid \eta > 2(p_h – (log(P_s)[I(t^*) – I(s)])) \\ &= \sigma_\eta [\varphi(\alpha/\sigma_\eta)/\Phi(\alpha/\sigma_\eta)] = \lambda, \end{aligned}$$

where $\alpha= 2(p_h – (log(P_s)[I(t^*) – I(s)]))$, φ is the standard normal density function, and Φ is the standard normal cumulative probability function. This expectation will non-zero for virtually every data point.

The impact of the error terms’ non-zero mean on the index estimate can be seen from the vector expression of the model. In vector terminology, the data generating process is given by

$$Y = XI + \Gamma + N \qquad (10)$$

where Y is a vector of log price differences, X is the matrix of repeat sales bases functions, Γ represents a vector of expectations of the error terms conditional on tiering, $\Gamma' = (\lambda_1,…, \lambda_T)$, one for each observation and N represents the vector of errors terms modified to have zero mean, i.e. where each element equals η - $E(\eta \mid tier = h)$. Applying least squares, the expectation of the resulting index estimate is given by

$$E(I) = I + (X'X)^{-1} X'\Gamma \qquad (11)$$

And so the bias is

$$b_0 = (X'X)^{-1}\ X'\Gamma = (X'X/T)^{-1}\ X'\Gamma/T \qquad (12)$$

This demonstrates there are two components, the inverse of the 'design' matrix X'X and the product of X and the vector of error expectations, Γ, that characterize the bias.

3. Bootstraping the Repeat Sales Model with House Price Tiers

The core idea of bootstrapping is that the properties of a statistic can be approximated through evaluating the distribution of the statistic calculated on a large number of resamples. In a range of applications, resampling has been shown to provide a useful approximation to the statistical properties under the true but unknown data generating process. There are a variety of ways of implementing the bootstrap. This paper utilizes the method referred to as 'error resampling' in which replications are created using draws of the sample residuals from an estimated model. In testing for the presence of tiers, we use a fully parametric bootstrap—where a parametric model is fit to the sample and resamples are created using the fitted model. This method is feasible because under the null hypothesis of no house price tiers, the residual from the fitted model are reasonably approximated by a normal distribution. In contrast, if house price tiers are a given, then the tiering methodology induces both a non-zero mean and non-normality for the residual distributions in each tier. In this case, the empirical residuals would need to be resampled after stratifying by holding period and tier.

To fix notation, define the true but unknown data distribution process (DGP) as F_0, the sample distribution as F_1, and the resampled distribution for replication r as F_{2r}. Similarly, the true but unknown house price index is given by $I_0 = (I^h_0, I^m_0, I^l_0)'$ and the hypothesis of no price tiers can be represented as the equality restriction, $I^h_0 = I^m_0 = I^l_0$.[6]

The estimated house price index based on the sample under the null (no price tiers) is denoted by $I_1 = I(F_1| H_o) = (I^*_1\ I^*_1\ I^*_1)'$. The house price index estimated under the alternative that allows for price tiers can be expressed as $I_1 = I(F_1) = (I^h_1\ I^m_1\ I^l_1)'$ and in this context the alternative hypothesis can be represented as $((I^h_1 - I^m_1)\ (I^m_1 - I^l_1))' \neq 0$ where the inequality holds strictly for at least one element of the vector.

The index estimation applied to the true parametric DGP, $\theta_0 = (I_0, \sigma_0(h))$, has a unknown bias, b_0, and results in a biased index estimate, I_1. The bootstrap estimate of bias, b_1, is based on the difference in the index used to create the bootstrap resamples, I_1, and the average index over the R bootstrap replications, $\{I_{21},\ldots, I_{2R}\}$:

$$b_1 = (\textstyle\sum_r I_{2r}/R) - I_1 \qquad (13)$$

[6] Note that by construction one element across the tiered repeat sales indexes will be exactly equal because they share a common normalization date. Also, note that the transpose of the index within parentheses is suppressed to avoid cluttered notation.

The bootstrap bias corrected index, I_{bc}, is then given by:

$$I_{bc} = I_1 - b_1. \tag{14}$$

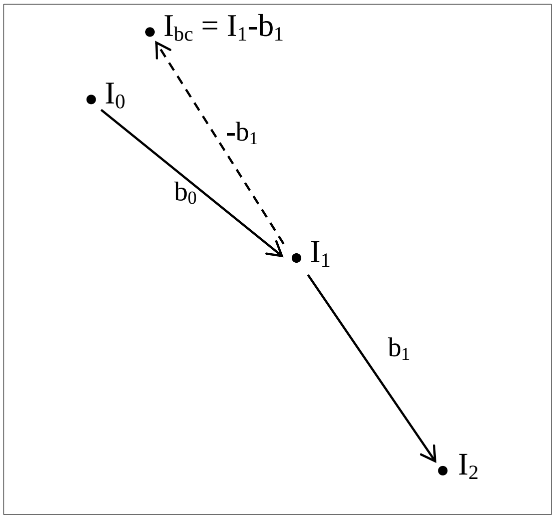


Figure 1: Bootstrap Bias Correction

Figure 1 illustrates the operation of the bootstrap bias correction. The bias of the tiered estimation results in the displacement of the estimated index, I_1, from the 'true' index, I_0, by an unknown bias, b_0. This unknown bias is approximated through the replications that treat I_1 as the 'true' index and estimate the bias as difference between the average tiered estimate, I_2 and I_1. Note that under the null hypothesis, estimating I_1 under the constraint that $I^h_1 = I^m_1 = I^l_1$, results in a consistent estimate of the 'true' unknown bias, b_0, because the constrained estimate is a consistent estimate of I_0.

To the extent that the bias corrected index I_{bc}, is a better estimate of I_0, than I_1, the estimation of the bias could be repeated starting at I_{bc} and generating a new bias estimate, b_2 and bias-corrected index, $I_{bc} = I_1 - b_2$. In principle, this algorithm could be iterated until the estimated bias function converges to a constant function.

4. Testing for the Existence of House Price Tiers

There are some differences between the approaches available for a statistical hypothesis test and that of estimation. First, under the null hypothesis of no price tiering, the indexes for the three possible tiers equal one another: $I^h = I^m = I^l$. Second, under the null, the data generation process is the 'standard' repeat sales model and hence it can be estimated using a single 'index' regression (with heteroskedastic errors). Let I^*_1 denote is this consistent estimator of I_0 estimated under the null.

The 'standard' index estimation applied to the sample is given by $I^*_1 = (I^*_1\ I^*_1\ I^*_1)'$ where the equality across tiers has been imposed in estimation. This constrained index is used to create the bootstrap resamples for each of the replications, but then tiered estimates are created using each replicate. For each replication, r, the observations are categorized into tiers and an estimate of the tiered index is made, $I_{2r} = (I^h_{2r}, I^m_{2r}, I^l_{2r})$.' Note that the estimate of each I_{2r} no longer constrains the index to be equal across tiers. The estimate of the bias is the difference in the average 'tiered' estimate of the index over the R bootstrap replications, $\{I_{21},\ldots, I_{2R}\}$, and the index used to create the replications, I^*_1:

$$b_1 = (\sum_r I_{2r}/R) - I^*_1. \tag{15}$$

Let the tiered estimation based on the original sample be denoted, $I_1 = (I^h_1, I^m_1, I^l_1)$. The bootstrap bias corrected index, I_{bc}, is then given by:

$$I_{bc} = (I^h_1, I^m_1, I^l_1) - b_1. \tag{16}$$

The variance/covariance matrix of the bias corrected tiered index, **Ω**, is estimated using the empirical variance/covariance matrix of the replicates. Specifically, this can be estimated by

$$\boldsymbol{\Omega} = \sum_r (I_{2r} - \sum_r I_{2r}/R)\ (I_{2r} - \sum_r I_{2r}/R)'/R, \tag{17}$$

or correcting for degrees of freedom,

$$\boldsymbol{\Omega}_{alt} = \sum_r (I_{2r} - \sum_r I_{2r}/R)\ (I_{2r} - \sum_r I_{2r}/R)'/(R - \text{length}(I) + 3). \tag{18}$$

The null hypothesis of no tiering can be expressed as a series of equality constraints on the estimated tiered index, namely $I^h_1 = I^m_1$ and $I^m_1 = I^l_1$. In vector notation, this constraint can be written $H' I = 0$ and the corresponding χ^2 given by $(H' I)' (H' \boldsymbol{\Omega} H)^{-1}(H' I)$ with degrees of freedom equal to 2/3 length(I) - 2.

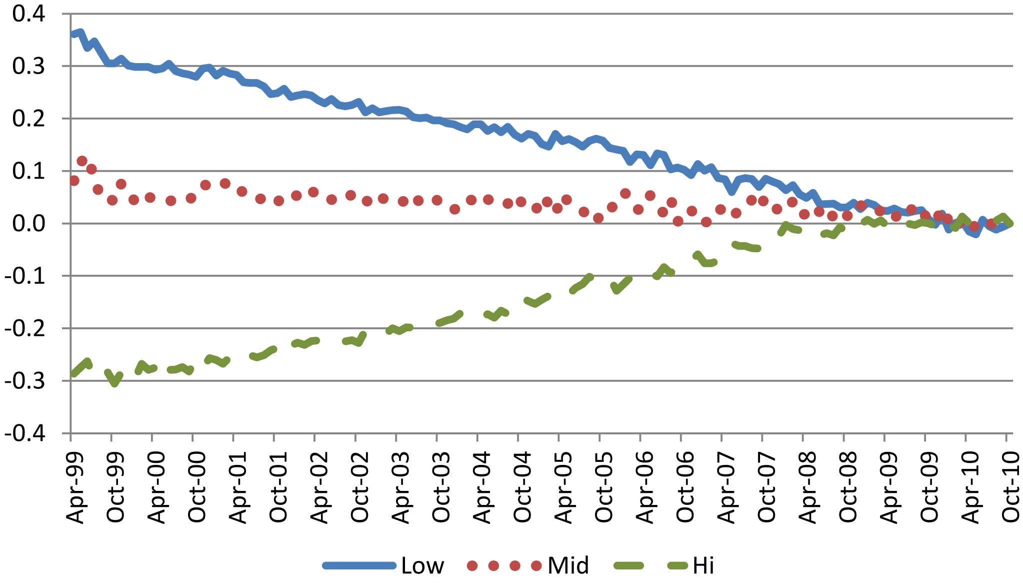
Table 1: Summary Statistics

	Sample Size (April 1999 - Oct 2010)	Mean	Std Dev	Min	Max
First Transaction Price (P1)	66,099	$264,704	$160,017	$18,500	$3,480,000
Second Transaction Price (P2)	66,099	$327,540	$194,498	$28,000	$3,200,000
Holding Period (months)	66,099	53	34	6	139
Annual Growth Rate	66,099	5%	15%	-37%	47%

5. Data

The data used to estimate the repeat sales model are based on home purchase loans originated between 1999 and 2010 in the state of California and purchased by Freddie Mac and Fannie Mae or available from a commercial data provider. Individual loans are matched into pairs based on property address. Exclusion filters were applied to the data in order to obtain a clean estimation dataset: First, transaction pairs with holding periods less than 6 months were excluded from the sample to avoid "flipped" properties; second, only properties with annual house price growth between the 5th and 95th percentiles were included to avoid contamination from outliers. The resulting data set includes more than 66,000 transaction pairs. Table 1 lists a few key statistics for the data.

Figure 2: Bias From Bootstrap Coefficient Estimates - California



6. Results

Figure 2 graphs the bias in the tiered repeat sales indexes for California estimated using 2000 bootstrap replications. It can be seen that the low tier index understates growth as this downward sloping bias function is subtracted from the raw low tier estimates (increasing the slope of the bias corrected index). This result matches intuition, as low value properties that have atypically large house price appreciation are likely to be re-classified as middle tier properties and leads to a downwards selection bias for the low tier segment. Conversely, the graph of the high tier bias function shows that the raw tiered index estimates will overstate the growth in this segment. Intuitively, middle value properties that have atypically large house price appreciation are likely to be re-classified as high tier properties, and high value properties that have atypically low house price appreciation are likely to be re-classified as middle tier properties, both leading the to an upwards selection bias in the high tier estimates. Note that the high and low tier bias functions are nearly symmetric across the horizontal axis. The middle tier bias function is approximately horizontal and so bias correcting has only minor impacts on estimated growth.

Table 2: Regression Approximation to Bias Correction

	CA		
	Low	Medium	High
Intercept	0.33702	0.06787	-0.31587
Date	-0.00255	-0.00044	0.00254
R^2	0.99020	0.60680	0.97870

A striking feature of Figure 2 is the apparent linearity of the bias function. Table 2 shows that the bias correction to the repeat sales index is very well approximated by a linear function of time. For each tier, the estimated bias at each date is regressed on the date (in months) and the results are reported above. For California, the linear approximation explains 99% and 98% of the variation in the bias function for the low and high tiers. The linear approximation only explains about 60 percent of the medium tier, but there is very little variation in the middle tier to explain, as it is nearly flat. Also, note that the linear coefficients for the high and low tiers are a very similar magnitude but of opposite sign, indicating a symmetry in the bias.

Figure 3: Aggregate vs. Bias-Adjusted Tiered Indices - California

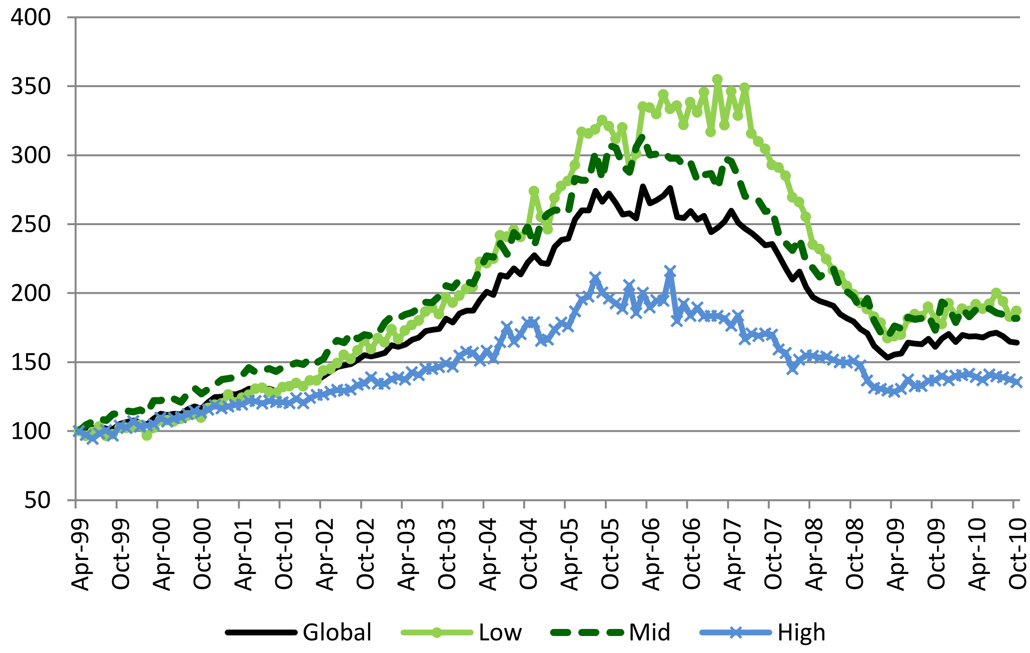


Figure 3 plots the aggregate and the bias-adjusted tiered indexes for California and displays that the bias correction acts to spread these estimated indexes apart. The differences in index across tier displayed in Figure 3 are statistically significant. Table 3 shows the chi-squared statistics for the test of equality of bias corrected indexes for the hypotheses that $I^h_1 = I^m_1$ and $I^m_1 = I^l_1$. A separate test statistic for $I^m_1 = I^l_1$ is given in the column labeled 'Low-Mid' and a test statistic for $I^h_1 = I^m_1$ is given in the column labeled 'Hi-Mid' and for these tests the critical value of the chi-squared statistic is 180. For California the chi-squared statistics for both these hypotheses is well above the critical value. Similarly, the results for joint test of $I^h_1 = I^m_1$ and $I^m_1 = I^l_1$ (which is null hypothesis of no price tiering) is strongly rejected at the 1% level having a critical value for the statistic of 360 and a chi-squared statistic value of 6068 for California. These chi-squared statistics for the test of equality of bias corrected indexes show that for California the null hypothesis of no price tiering is strongly rejected at the 1% level.

Table 3: Chi-Square Statistic: Test of Presence of Bias Effect

	Low-Mid	Hi-Mid	All
State: CA	**1418**	**1954**	**6068**
Critical Value at 1% Level	180	180	360
# Restriction /Degree of freedom	138	138	276

7. Conclusions

Understanding the forces that guide house price movements is important for homeowners, real estate professionals, and the mortgage finance industry. Practitioners have long identified price tiers in housing markets as an important factor in understanding local housing markets. In the current recovery, for example, there may be some markets with uneven appreciation across price tiers, as mortgage credit availability continues to be especially restricted in high-value tiers. However, there have not been valid methods of testing for and estimating tier effects.

This paper proposes a new methodology to address this gap in the contest of repeat sales index estimation. This method is applied to a set of California repeat sales transactions, and demonstrates the existence of material tier effects in home sales paths. While this tool was implemented in a simple case of three market tiers, it generalizes to any number of tiers.

References

Bailey, M. J., R. F. Muth, and H. O. Nourse. 1963. A regression model for real estate price index construction. Journal of the American Statistical Association 58:933–942.

Case, K.E., and R.J. Shiller 1987. Prices of single-family homes since 1970: new indexes for four cities. New England Economic Review, Federal Reserve Bank of Boston, 45-56.

Case, K. E., and R. J. Shiller. 1989. The efficiency of the market for single-family homes. American Economic Review 79:125–137.

Delaney, C.J., J.A. Seward, and M.T. Smith. 1992. An empirical analysis of housing price appreciation in a market stratified by size and value of the housing stock. The Journal of Real Estate Research, vol. 7, no. 2 (Spring), pp. 195-205.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Annals of Statistics vol. 7, pp. 1-26.

Gatzlaff, D., and D. R. Haurin. 1997. Sample selection bias and repeat-sales index estimates. Journal of Real Estate Finance and Economics, 14: 33-50.

Goodman, A. (1978) Hedonic prices, price indices, and housing markets," Journal of Urban Economics 4 (1978): 471-484

Hall, P. 1992. The bootstrap and edgeworth expansion. Springer-Verlag New York.

Haurin, D. R. and P. H. Hendershott. 1991. House price indexes: Issues and results. AREUEA Journal, vol. 19, no. 3 (Fall), pp. 259-69.

Heckman, J. 1979. Sample selection bias as a specification error." Econometrica, vol. 47, pp. 153-61.

Horowitz, J. 2001. The bootstrap. Handbook of Econometrics, edited by J.J. Heckman and E. Leamer, volume 5, chapter 52, 3159–3228.

Leventis, A. 2012. Home price indexes for homes in different price tiers: Biases and corrections. Federal Housing Finance Agency Working Paper 12-1

Mayer, C. 1993. Taxes, income distribution, and the real estate cycle: Why all houses do not appreciate at the same rate. New England Economic Review, May/June 1993, available at: www.bos.frb.org/economic/neer/neer1993/neer393c.pdf.

Pollakowski, H. O., M. A. Stegman, and W. Rohe. 1991. Rates of return on housing of low- and moderate- income owners. AREUEA Journal, vol. 19, no. 3, pp. 417-25.

Poterba, J. M. 1984. Tax subsidies to owner-occupied housing: An asset-market approach." The Quarterly Journal of Economics (November), pp. 729-52.

Poterba, J. M. 1991. House price dynamics: The role of tax policy and demography. Brookings Papers on Economic Activity 2, pp. 143-83.

Wang, F. T., and P. M. Zorn. 1997. Estimating house price growth with repeat sales data: What’s the aim of the game? Journal of Housing Economics 6(2), 93–118.