

Home Price Index

Joshua Roberson
April 3, 2019

Publication 2230



Over the past few years, Texas has led U.S. states in both economic and population growth. It's tough to find another state that matches or exceeds the Lone Star State in consecutive year-over-year growth. With so much growth, housing needs have changed dramatically, and home prices and affordability have become hot topics.

For many years, Texas boasted spacious and affordable housing. Compared with other large-economy states such as California and New York, Texas can still make these claims. However, by Texas standards housing affordability may have changed for good.

One of the more interesting Texas home-price trends has been how widespread price growth has been for a state so large. Austin, Dallas-Fort Worth, Houston, and San Antonio have led the charge not only in price growth but also in home scarcity and marketing time. Many mid-major metros have also shared in rapid price growth, including Lubbock, Tyler, and Waco. Job creation and population growth have spurred housing markets across the state.

Keeping an Eye on Price Growth

Several options are available to track price changes, which may sound surprising for a nondisclosure state.

One of the better-known resources is monthly aggregates. There are several outlets for these aggregates, including local Realtor boards, Texas Realtors, and the Real Estate Center. These aggregates provide simple and intuitive measures to track home prices. However, their usefulness for tracking price growth is limited due to differing sample characteristics from one period to another. For example, the sample of homes sold in a particular market last June may not resemble those from June ten years ago. As trends, styles, supply, and demand change, what ends up getting analyzed may simply be the change in marketing mix available. While such an analysis is telling, it does not sufficiently answer the question of market home-price growth.

Repeat sales provide an alternative approach more tailored to home-price appreciation. Unlike the simpler aggregates previously mentioned, the repeat-sales

approach takes additional measures to filter for homes that have sold at least twice. This means an even smaller final sampling size once grouped into sale pairs, which normally limits this approach's usability to larger markets. This data arrangement allows for price comparisons on the same houses over time. Also, repeat sales are held to a constant quality constraint that eliminates homes with known improvements. These steps better address the marketing mix problem and are better suited to isolate market price growth.

Price-trend modeling with repeat sales data was popularized in the Case-Shiller model. This model is used by entities such as the Federal Housing Finance Agency (FHFA), Freddie Mac, U.S. Dow Jones Indices, and now the Center. There are two major variations of the Case-Shiller model. The original variant is a geometric model that measures percent price growth, and the latter is an arithmetic model that measures absolute price change. Further differences between the two are covered later.

Data Preparation: Sale Pairs

Before any modeling can occur, data must first be prepared from raw sales to repeat sales then into sale-pair groups. Typically, sales data will be available in a tabular format (Table 1).

Table 1. Tabular Sales Data

Property Identifier	Sale Date	Sale Price
A	Q1	\$100,000
B	Q2	\$110,000
A	Q2	\$120,000
A	Q3	\$130,000

Source: Real Estate Center at Texas A&M University

Each transaction is given its own row or record with sale-specific information. In this case, property A sold three times within the series timeframe. Transforming sales data into repeat sales is as simple as filtering for properties that have sold more than once. This is a relatively straightforward step provided you have reliable property identifiers. Given a large enough series of repeat transactions, sale-pair data can be produced for a given market. This is where most data shrinkage occurs. Table 2 provides a good model of what the end transformation to sale-pair groups looks like.

Table 2. Tabular Sale-Pair Data

Property Identifier	Begin Sale Date	Begin Sale Price	End Sale Date	End Sale Price
A	Q1	Q2	\$100,000	\$120,000
A	Q2	Q3	\$120,000	\$130,000

Source: Real Estate Center at Texas A&M University

For property A, this data transformation will eventually shift attention from house price in period X to the price change from period X to Y. This format is more conducive to tracking and measuring price appreciation.

In addition to data reformatting, steps are taken to eliminate sales that may be influenced by nonmarket forces. Such sales include non-arms-length transactions (those between unrelated parties acting in their own best interest), home flips, and properties with significant home improvements. Common filters include eliminating properties with explicit transaction flags, obvious data-entry errors, exceptionally short holding periods, or noticeable changes in square footage. Identifying and establishing rules for these factors is much easier said than done and could be described as more art than science.

In the end, the processed sale-pair data will represent a much smaller percentage of the original data set. This is one of the major critiques against repeat-sale methodology. At the time of this writing, the Real Estate Center had over 1.6 million sales for the Dallas-Fort Worth metropolitan area, including almost 750,000 identified repeat sales. After transforming and filtering the records, just over 250,000 paired sales remained, a fraction of the original data set.

This highlights what is probably the biggest limitation of using sale pairs for modeling. The process appears wasteful of data and limits analysis to relatively large housing markets. In addition, filtered data is often viewed as less representative of the overall market largely because of the elimination of new-home sales.

Case-Shiller Model

While the names Case and Shiller are most synonymous with home price indexes, their model was an improvement on another model developed by Bailey, Muth, and Nourse (BMN). In 1960, BMN developed a model that estimated coefficients for quarter-to-quarter market growth.

$$[1] \frac{P_{it'}}{P_{it}} = \frac{B_{t'}}{B_t} \epsilon_{itt'}$$

In this model, both t' and t represent two separate periods for each sale pair—the later period, t' , and the earlier period, t . Here, the price growth between sale pairs from

one period to another ($P_{it'}/P_{it}$) is estimated by the market growth rate between the same two periods ($B_{t'}/B_t$) times an error term.

$$[2] P_{it'} - P_{it} = b_{t'} - b_t + \epsilon_{itt'}$$

The model above provides an alternate view in logarithmic form where p_{it} represents the log price from the first transaction in a sale pair for a particular property while $p_{it'}$ represents the log price for the second transaction. Market growth rate coefficients are regressed using dummy variables as independent variables. Dummy variables fill a sparse matrix where rows are made up of individual sale pairs and columns are individual periods. For each sale pair a -1 is placed in the period when the first transaction occurred and a 1 for when the second transaction occurred. All other matrix elements are zero. The example below shows a sample of sale pairs in matrix form.

[3] Sale pairs in matrix form

$$Y = \begin{bmatrix} P_{12} - P_{11} \\ P_{22} - P_{21} \end{bmatrix} \quad X = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Here, the first row represents an individual sale pair purchased in the first period of the series and sold in the following period. The second row represents another sale pair that was bought in the first period and sold in the third. In matrix X, the elements in the first column that are not zero represent transactions occurring within the first period within the index series.

From here, least-squares regression would be used to estimate coefficients relating Y, or the change in value, to individual periods. These coefficients are still in logarithmic form and need to be transformed exponentially to produce model index values.

$$[4] \hat{\beta} = (XX')^{-1} (XY)$$

The error term for BMN's model is assumed to be homoscedastic and not related between one sale pair or another. Intuitively, the error term represents transaction-level pricing misinformation that is independent from other transactions.

In the late 1980s, Case and Shiller amended the original BMN model by assuming that the error terms were heteroskedastic based on properties with longer holding periods. Their intuition was that properties with longer hold periods contained price deviations besides normal price misinformation. To address this issue and produce a better, unbiased model, they added holding period weights regressed from the error terms from BMN's original model.

The error term from the first regression could be further decomposed into the following:

$$[5] \epsilon_{itt'} = m_{itt'} + h_{itt'}$$

The first error component is measured as the pricing error due to transaction participant misinformation and/or rash decision-making like buying too quickly. This component has the statistical characteristics of a white-noise process where $m \sim \text{Normal}(0, \sigma_m^2)$. The second error component consists of a random walk-process interval error that represents differences in market tastes at a given period. The difference in interval errors also makes up a normal distribution where $\Delta h \sim \text{Normal}(0, \sigma_h^2)$ and is estimated by an individual sale pair's holding period. These estimators can be produced through the following second-stage model.

$$[6] \hat{\epsilon}_i^2 = 2\sigma + 2(t' - t) + n_{itt'}$$

This final model accounts for the sale pair influence with longer holding periods. Here $(t' - t)$ represents the holding period between sales in each sale pair. Fitted values from this model are used to produce weights in a third and final weighted regression model.

Case-Shiller Model Variants

The model previously outlined was a geometric model and is essentially what is used by the FHFA and Freddie Mac for their various quarterly home price indices. An alternative arithmetic model was later developed by Robert Shiller and is used by U.S. Dow Jones Indices to produce their S&P Core Logic Case-Shiller Index.

One of the biggest differences between the geometric and arithmetic models is that, by design, the arithmetic model is more influenced by higher priced homes. In fact, the index produced by the arithmetic model is analogous to a capital-weighted index such as the NASDAQ Composite. The geometric model, on the other hand, weights each sale pair equally.

The arithmetic model also differs by using an instrumental variable regression. This is because the model contains actual observed bought and sold prices. These observed prices are assumed to contain data entry errors and, therefore, correlate to the error term. To account for this correlation, which violates the Gauss-Markov theorem, instrument dummy variables are used. The dummy variable structure used is essentially identical to the geometric model where -1 is placed for a purchase, 1 is placed for a sale, and zeros for everything else.

$$[7] \quad Y = \begin{bmatrix} P_{10} \\ P_{20} \\ P_{30} \\ P_{40} \\ 0 \end{bmatrix} \quad X = \begin{bmatrix} P_{11} & 0 \\ P_{21} & 0 \\ 0 & P_{32} \\ 0 & P_{42} \\ -P_{51} & P_{52} \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ -1 & 1 \end{bmatrix}$$

The first row represents a sale pair purchased in the base period and sold in the first period. Sale pairs with an initial purchase within the base period are not given a negative dummy variable, but all other initial purchases are. Also, each value of X has a matching Z instrumental variable dummy variable.

Final Model Selection

Both the geometric and arithmetic models were evaluated side by side for various Texas markets. Ultimately, the Center adopted the geometric model due to its equal weighting of each sale pair. As discussed in the previous section, the design of the arithmetic model produces an index weighted heavier on higher-priced homes.

Given the nature of the Center's housing data, which include many high-priced homes, the indexes produced by the geometric model proved less volatile than the arithmetic model. For example, a multimillion-dollar home with a 1 percent change in value has a tremendous impact in the arithmetic model. The absolute change in value for this home can have a disproportionate effect on the model results.

Model Comparison

Currently, Case-Shiller models are used by entities such as the FHFA, Freddie Mac, U.S. Dow Jones Indices, and the Real Estate Center for publicly available home price indices. In its All-Transactions Index, the FHFA uses pricing data available from mortgage application data and accompanied appraisals from both Fannie Mae and Freddie Mac. Because of the widespread use of financing for residential real estate reporting, coverage in Texas is fairly complete for many Texas markets. Besides the All-Transactions Index, FHFA also produces a Purchase-Only Index that is most similar to the Center's index based on the target sample of single-family purchases.

U.S. Dow Jones Indices publishes the S&P CoreLogic Case-Shiller

Index, which is a monthly arithmetic Case-Shiller index.

Through the Center's participation in the Data Relevance Project, a research agreement with Texas Realtors, repeat sale indices can now be produced for a number of Texas markets (see Figures 1–6).

The primary benefit of the Center's Home Price Index is the dataset used for modeling. Compared side by side with the FHFA index, also a geometric model, the Center's index shares a similar trend but has a flatter curve. The difference appears to be due to the Center's broader distribution of market price points compared with the FHFA. The FHFA is limited to homes financed using conforming conventional loans, currently capped at \$453,100 in Texas. The Center's index includes all single-family home sales reported through the local MLSs.

One of the more interesting types of analysis is comparing index trends by price tiers. Figure 7 illustrates the differing price-growth rates for three separate price tiers in Austin-Round Rock. Price-change trends in the highest price tier appear to be less aggressive in magnitude compared with lower price tiers. Growth-rate patterns are apparent in both magnitude and volatility among the three groups. The geometric model's equal weighting allows for the better capturing of comprehensive market-price growth.

Index Availability

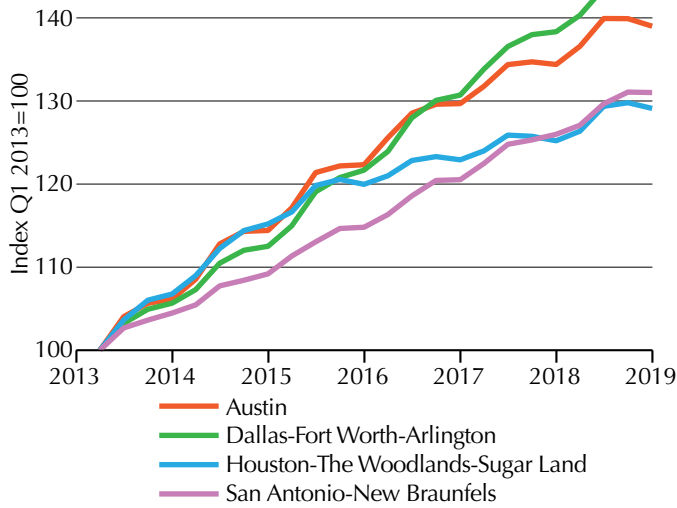
The criteria for potential markets for home price index production depends primarily on the availability of identifiable repeat sales across a timeline with multiple market cycles. Quarterly indices will be produced for the four major Texas metros as well as several mid-sized metros (Table 3). Monthly indexes will be produced for the four major metros.

Table 3. Real Estate Center Home Price Index Metros

Metro	Base Year	Frequency
Amarillo	2002	Quarterly
Austin-Round Rock	1999	Monthly, Quarterly
Dallas-Fort Worth-Arlington	2005	Monthly, Quarterly
Dallas-Plano-Irving	2005	Monthly, Quarterly
El Paso	2004	Quarterly
Fort Worth-Arlington	2005	Monthly, Quarterly
Houston-The Woodlands-Sugar Land	2000	Monthly, Quarterly
San Antonio-New Braunfels	2013	Monthly, Quarterly
Sherman-Denison	2014	Quarterly

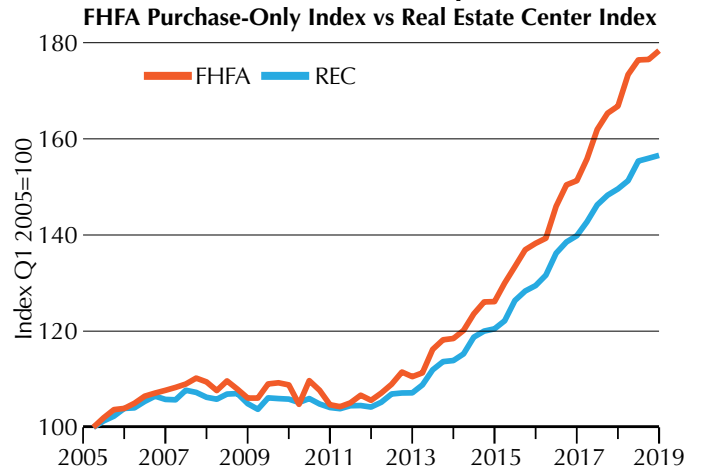
Source: Real Estate Center at Texas A&M University

Figure 1. Texas Metro Index Comparison



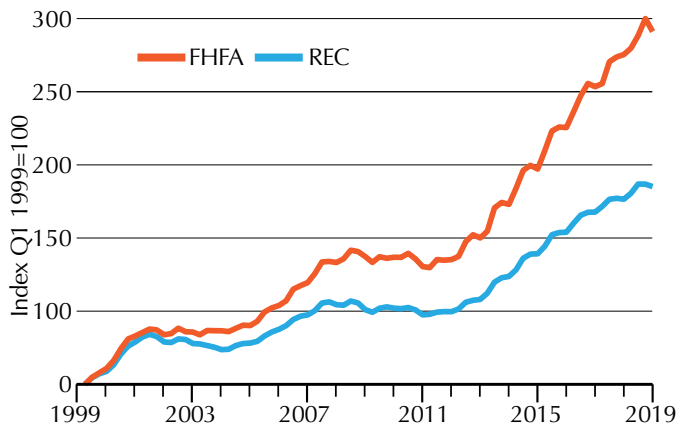
Source: Real Estate Center at Texas A&M University

Figure 4. Fort Worth-Arlington Home Price Index Comparisons



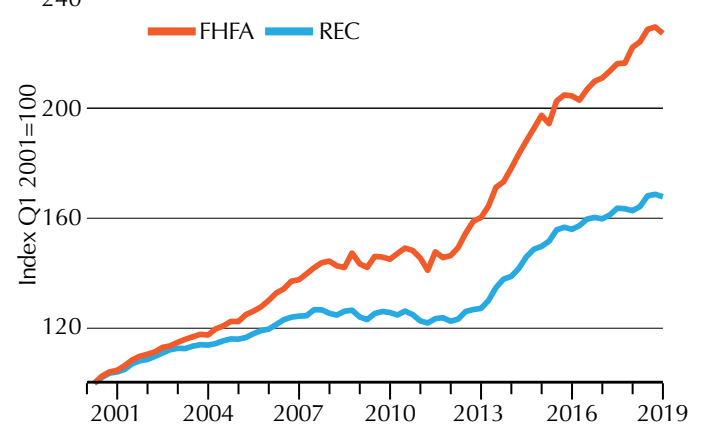
Sources: Federal Housing Finance Agency, Haver Analytics, and Real Estate Center at Texas A&M University

Figure 2. Austin-Round Rock Home Price Index Comparisons
 FHFA Purchase-Only Index vs Real Estate Center Index



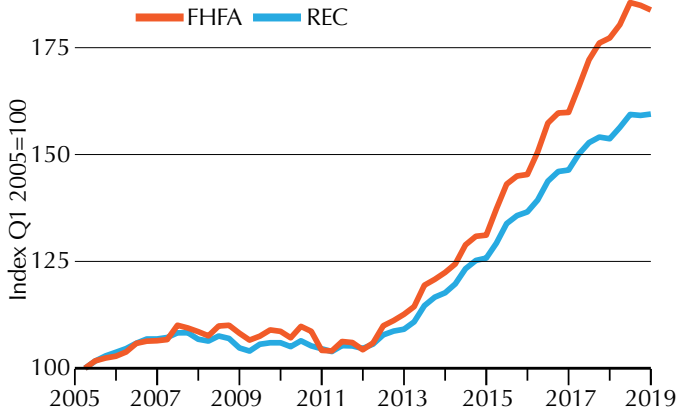
Sources: Federal Housing Finance Agency, Haver Analytics, and Real Estate Center at Texas A&M University

Figure 5. Houston-The Woodlands-Sugar Land Home Price Index Comparisons
 FHFA Purchase-Only Index vs Real Estate Center Index



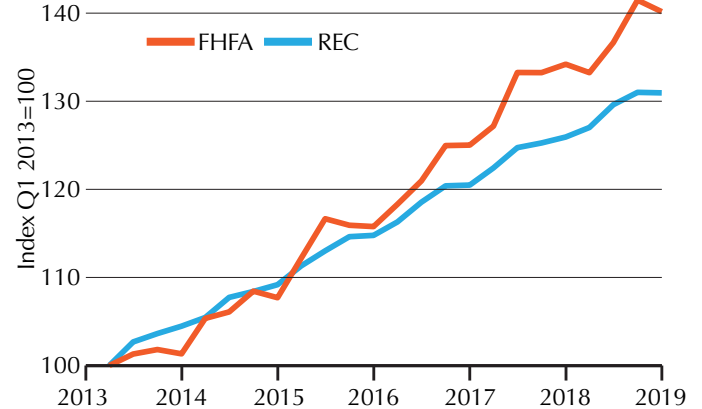
Sources: Federal Housing Finance Agency, Haver Analytics, and Real Estate Center at Texas A&M University

Figure 3. Dallas-Plano-Irving Home Price Index Comparisons
 FHFA Purchase-Only Index vs Real Estate Center Index



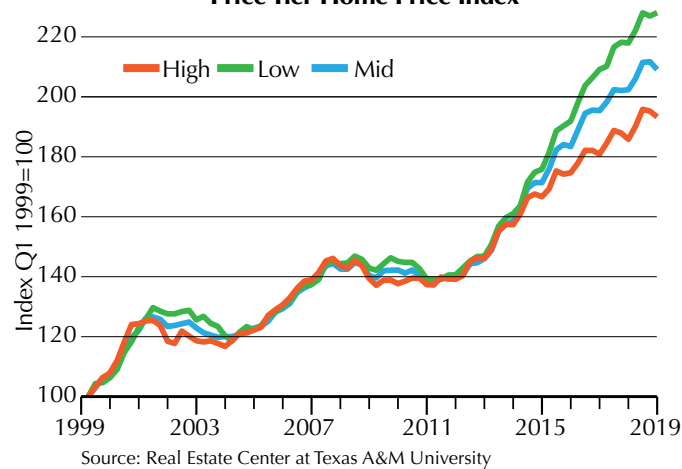
Sources: Federal Housing Finance Agency, Haver Analytics, and Real Estate Center at Texas A&M University

Figure 6. San Antonio-New Braunfels Home Price Index Comparisons
 FHFA Purchase-Only Index vs Real Estate Center Index



Sources: Federal Housing Finance Agency, Haver Analytics, and Real Estate Center at Texas A&M University

**Figure 7. Austin-Round Rock
Price Tier Home Price Index**



References

Bailey, Martin J.; Muth, Richard F.; Nourse, Hugh O. "A Regression Method for Real Estate Price Index Construction," *Journal of the American Statistical Association*, Vol. 58, No. 304 (Dec., 1963), pp. 933-942

Shiller, Robert J. "Arithmetic Repeat Sales Price Estimators," *Journal of Housing Economics* 1, 110-126 (1991)

Office of Federal Housing Enterprise Oversight. OFHEO House Price Indexes: HPI Technical Description. <https://www.fhfa.gov/PolicyProgramsResearch/Research/Pages/HPI-Technical-Description.aspx>

S&P Dow Jones Indices. S&P CoreLogic Case-Shiller Home Price Indices Methodology. https://us.spindices.com/documents/methodologies/methodology-sp-corelogic-cs-home-price-indices.pdf?force_download=true

National Bureau of Economic Research "Prices of Single-Family Homes Since 1970: New Indexes for Four Cities," Working Paper No. 2393

© 2019. Real Estate Center. All rights reserved.

