StreetEasy Condo Market Index for Manhattan
Index Construction Methodology

Sam Lin, Sofia Song, & Sebastian Delmont

Introduction

We are pleased to present the StreetEasy Condo Market Index (CMI) for Manhattan as a unique guide for tracking price movements in the Manhattan real estate market.

The Manhattan real estate market is like no other housing market in the country. For an island that comprises only 14 square miles, Manhattan is the most watched real estate market in the world. While the Case Shiller Housing Index is often touted as the barometer of the country's real estate market, many will argue that it has very little relevance to the Manhattan market as it only includes single-family homes and covers such a wide-metro area ranging from Pennsylvania to Connecticut. In contrast, the Manhattan market is almost entirely composed of condo and co-operative apartment sales.

StreetEasy took on the task of creating an index strictly for the Manhattan condo market in order to provide a useful and practical tool in gauging the health of the Manhattan real estate market. This paper will outline the methodology employed in constructing this index.

Why not include co-ops?

For the purposes of our index and study, we excluded co-op sales. Condos are real property. They are tangible and their values are easily measured. In contrast, purchasing a co-op means purchasing shares in a corporation. Co-ops are jointly owned among all the shareholders and therefore, the values of individual units are difficult to separate from the value of the entire corporation. Additionally, the data currently available is much more reliable for condos than it is for co-ops. It was not until 2006 when the NYC Department of Finance required that co-op sales disclose the full purchase price. Even still, obtaining unit numbers for co-op units have proven unreliable and difficult. In terms of market share, condos made up 50.2% of all apartment closings in Manhattan, while co-ops made up 49.8%, from January 2006 through August 2010.

Why not just use Average Price Per Square Foot (PPSF)?

A 'quick and dirty' approach in tracking price movement is to just take the average PPSF of one period and compare it to another period. However, this simple approach has several disadvantages and may not accurately measure the health of the Manhattan real estate market. The mix of properties sold in one time period may not necessarily be comparable to the mix of properties sold in a subsequent period; they may be entirely different in terms of age, size, quality, etc. For example, as seen in this past decade during the real estate boom, inventory was tight and buyers were buying whatever units they could get, even if the quality was significantly less than desirable. In contrast, in a down market like that of the last two years, buyers had their pick of properties and in general, would choose to purchase property that would be of the best

quality, and would present the lowest risk in investment. As a result, the average PPSF approach can be very much like comparing "apples to oranges."

A repeat sales transaction-based index allows for an "apples to apples" approach and is more like a stock market index as it tracks price changes of the same properties (or in the case of the stock market, the same stock) over time. Since this approach compares literally the same properties, errors or biases created by variables like location, size, age, and quality, are minimized.

Index Construction & Model Formulation

Data Source

We used recorded sales data provided by the New York City Department of Finance from January 1, 1995 through August 2010. Since condos are real property, they are each assigned their own borough, block and lot number (BBL). We were able to identify repeat sales by matching: property type (condos only), address, and BBL among the 101,775 records we collected. From these records, we used the following information:

- Address, street name, building name, unit number, and neighborhood
- Final sales prices in $US
- Transaction date (the date the property was legally transferred to the new owner)

Frequency of Sales for Condos in Manhattan, January 1995 – August 2010

# of Paired Sales	# of Properties	# of Transactions
6	6	42
5	22	132
4	194	970
3	1,128	4,512
2	5,158	15,474
1	18,783	37,566
Total	25,291	58,696

Step 1: Using a Base Model: Standard Repeat Sales Regression (RSR)

In developing this index, we first did a survey of indices that measured the apartment market, such as in Tokyo and Hong Kong, and studied their methodologies. We also analyzed other housing indices such as the Case Schiller Index and the Conventional Mortgage Home Price Index, developed by Fannie and Freddie Mac.

Like other housing indices, we started with a transactions-based approach and measured price changes since 1995 by using the repeated sales methodology.[1] Since each residential property has its own unique features and characteristics, such as location, size, views, amenities, and

[1] A repeat sale can be defined as at least one pair of transactions for the same property in the 15 years' worth of closing data we have, from January 1995 through August 2010.

condition, the price of each property is therefore quite different from one another and would result in an "apples to oranges" comparison. Therefore, in order to create an "apples to apples" or like-to-like comparison, we used data for condo units that sold at least twice, i.e., had repeat sales over this 15-year period, to determine the price change for the same unit assuming the quality and size of each condo would remain constant. Using this criterion, we were able to use 58,696 repeat sales records for 25,291 different properties, out of 101,775 sales transactions.

The starting point for our index was the standard repeat sales regression (RSR) analysis, a widely used statistical method developed in 1963 by Bailey, Muth and Nourse.[2]. The RSR is a statistical way for estimating price change (what we call the dependent variable), which is based on the value of other variables such as time (this is what we would consider to be an independent variable). Essentially, the model uses rates of return between when a property sells in one time point and when it sells in a subsequent time point.

The RSR model starts with the assumption that characteristics and qualities of the properties remain the same between each pair of transactions, and that the only thing changing is the price. The model basically says that the price paid for a property is the value of all the different attributes (like location, size, age, quality) that is unique to this property. For example, if we know how much a condo sold in one year and we know how much it sold for three years later, we will therefore know how much the value of this condo changed over time.

Each paired sale is known as an observation. Each observation consists of:

1. the first sales price of property *i*
2. the time period of the first sale
3. the second sales price, and
4. the time period of the second sale.

$$\ln\left(\frac{P_{i,t2}}{P_{i,t1}}\right) = \sum_{t} a_t D_{i,T} + \ln \varepsilon_i$$

Here in this regression equation, the log of the ratio of the second sales price to the first sales price, of each repeat sales pair, is the dependent variable (the left-hand side of the equation). The value of this ratio depends on the right-hand side of the equation, the independent variables, which are comprised of a_t multiplied by a dummy variable and the log of an error term.

$P_{i,t1}$ is the price of property i, at the first sale, in time point 1
$P_{i,t2}$ is the price of property i, at the second sale, in time point 2
a_t is a coefficient or parameter that is estimated and determines the index value
$D_{i,T}$ is an independent dummy variable for a series of months (between January 1995 and August 2010) where:

[2] Bailey, M.J., R. F. Muth, and H.O. Nourse, 1963, "A Regression Model for Real Estate Price Index Construction," Journal of the American Stratistical Association 58, 933-942.

$$D_{i,T}\begin{cases} -1 & \textit{when the initial sale took place} \\ +1 & \textit{when the second sale took place} \\ 0 & \textit{when no sale has take place} \end{cases}$$

ε_i is an error term

The resulting index is created by converting the time coefficients, a_t, into index values.

Step 2: Adjusting the RSR using Case-Shiller's methodology

One problem with the RSR base model is that it does not completely account for the actual quality of the property. It can be assumed that the houses that have had recent renovations and improvements will sell more quickly, and therefore, the RSR data sample will be dominated by these properties. Other properties that take a long time to sell, and would therefore have longer time intervals may be of poorer quality. The Case-Shiller Index (CSI) tries to control for these biases by using a weighted least squares technique where the weights are inversely related to the length of time between sales, e.g., weighting transactions with longer time intervals less than transactions that had shorter time intervals. Additionally, the CSI excluded data like new developments, condos, non-market transactions (e.g., between family members), and sales that took place in less than three months.

An additional issue with the RSR is that errors were different for each paired sale or heteroscedastic. Case and Shiller adjusted this problem in a method they proposed in 1987.[3] In the RSR model, errors are likely to become larger when properties are sold after long time intervals, which have a larger influence on the index relative to properties sold over short time intervals. Case-Shiller assumed that errors came from two sources:

1. Time intervals between paired sales. There are errors that arise in repeat sale pairs due to the varied lengths of time between transactions for different properties. It was found that the longer the time interval, the larger the pricing errors. (Since errors varied for every paired sale, this is known as heteroscedasticity.)
2. Mispricing errors, which are likely to be independent and not correlated.

The error terms can be expressed as:[4]

$$\boldsymbol{E_{i,t} = h_{i,t} + m_i}$$

$\boldsymbol{h_{i,t}}$ is the transaction interval error for pair i at time t;

$\boldsymbol{m_i}$ is the mispricing error which is the variance of the property-specific random error.

The mispricing errors are likely to be independent, both across pairs of sales, as well as, transaction periods, and can be represented by an identically distributed white-noise term:

---

[3] Case, E. Karl, and Robert J. Shiller, 1987, "Prices of Single Family Homes Since 1970: New Indexes for Four Cities," *New England Economic Review*, Sept./Oct. pp 60-62

[4] S&P/Case-Shiller Metro Area Home Price Indices Index Methodology, p17. http://www2.standardandpoors.com/spf/pdf/index/SPCS_MetroArea_HomePrices_Methodology.pdf

$m \sim \text{Normal}(0, \sigma^2_m)$ is the variance of the mispricing errors.

In contrast, the transaction interval errors are assumed to follow a Guassian random walk**,** so $\blacktriangle h \sim \text{Normal}(0, \sigma^2_h)$ and the longer the length of the time between sales, the larger the variance of the interval error.

As a result, the variance of the combined mispricing and interval errors for any paired sale is likely to be assumed as: $\mathbf{2\sigma^2_m + I_i \sigma^2_h}$ where I_i is the time interval between sales for pair i.

To adjust for heteroscedasticity, Case-Shiller adjusted the model in three ways:
1. They estimated repeat sales parameters using the Ordinary Least Squares (OLS) method
2. They regressed the squared residuals obtained from the first stage on a constant and on the length of time between sales
3. They re-estimated the repeat sales parameters using Weighted Least Squares (WLS) where the weights are inversely proportional to the predicted residuals obtained in the second stage.

Step 3: Fine-tuning the CSI and the RSR for StreetEasy's Condo Market Index (CMI)
Building upon Bailey-Muth-Nourse's RSR and the CSI, the StreetEasy (CMI) made the following assumptions in our regression analysis:
1. The data sample used is representative;
2. The variables are error free;
3. The errors are uncorrelated;
4. The variance of error is homoscedastic, or constant across observations.

However, the real world is not perfect and there are several factors that create errors and biases, and skew pricing trends. To minimize these errors, we adjusted the following:

1. Data size: The sample data size consisting of the number of paired sales, over some time periods were small, lagged, or highly fluctuating. We used a three-month moving average to estimate the trend line, similar to the CSI, to adjust for these limitations and to reduce statistical errors. The sample size of each month is comprised of the reporting month and preceding two months. For instance, the May index is based on the samples of repeated sales in May, April, and March.

2. Data filters: We excluded the following data:
    a. Price outliers – extremely high or low prices, as large price changes indicate that:
        i. there may have been data entry errors, or
        ii. the properties may have had significant changes to their quality, or
        iii. there may have been transfers between family members/friends (non-arm's length);
    b. Foreclosures, transfers between family members, and income-restricted sales were excluded as they do not reflect true market value;
    c. Time share and hotel condos were also omitted as they are considered to be a different market;

d. Although new development sponsor sales were included, they were not included in the index until they had a repeat sale, to be included in a *pair* of transactions.

While the CSI excludes properties with high turn over – repeat sales in less than six months' time - as these sales are viewed to be 'flips' and would skew the price higher, the StreetEasy CMI includes these transactions as they are considered to be part of the common behavior found in the Manhattan real estate market. The Manhattan market is unique in that holding periods of properties are rather short (an average of three years) and there are properties with a high turnover, producing a healthy sample size of 25,045 repeat sales in 15 years.

We scaled our index by assigning January 2000 as the base period with a value of 1000.

Considerations

Timeliness of Data

While this index is similar to a stock market index as it tracks the same properties, the biggest difference is the timeliness of the data. A stock market index will be adjusted every day with up-to-the-minute pricing. However, due to the nature of the real estate market and the difficulty in closing a sale, the data for the index is not that timely and suffers from two sources of lag:

1. The data used here are public records where sales closings were recorded with the New York City Department of Finance. Typically, in the Manhattan condo market, it takes two to three months from the contract signing to close the transaction. However, the market climate can change dramatically from the time a contract is signed until it actually closes. By the time a unit sells and closes, the price is actually about three months old and reflects the market conditions of when it went into contract.
2. The other source of lag is the recorded date – the date the sale actually gets registered with the New York City of Department of Finance. While New York City requires that all sales be recorded with the City within 15 days of the closings, we found that most sales get recorded within two to eight weeks of the closing date. While StreetEasy receives new data from the Department of Finance every day, the price data received could often be a total three to five months old.

Sample Selection Bias

It can be argued that the use of repeat sales data is not really a random, representative sample of the market in the short run. The argument is that properties, which are often sold multiple times, differ in attributes and quality than those that are seldom sold. For instance, entry-level properties have more frequent transactions than larger, higher-priced properties.[5]

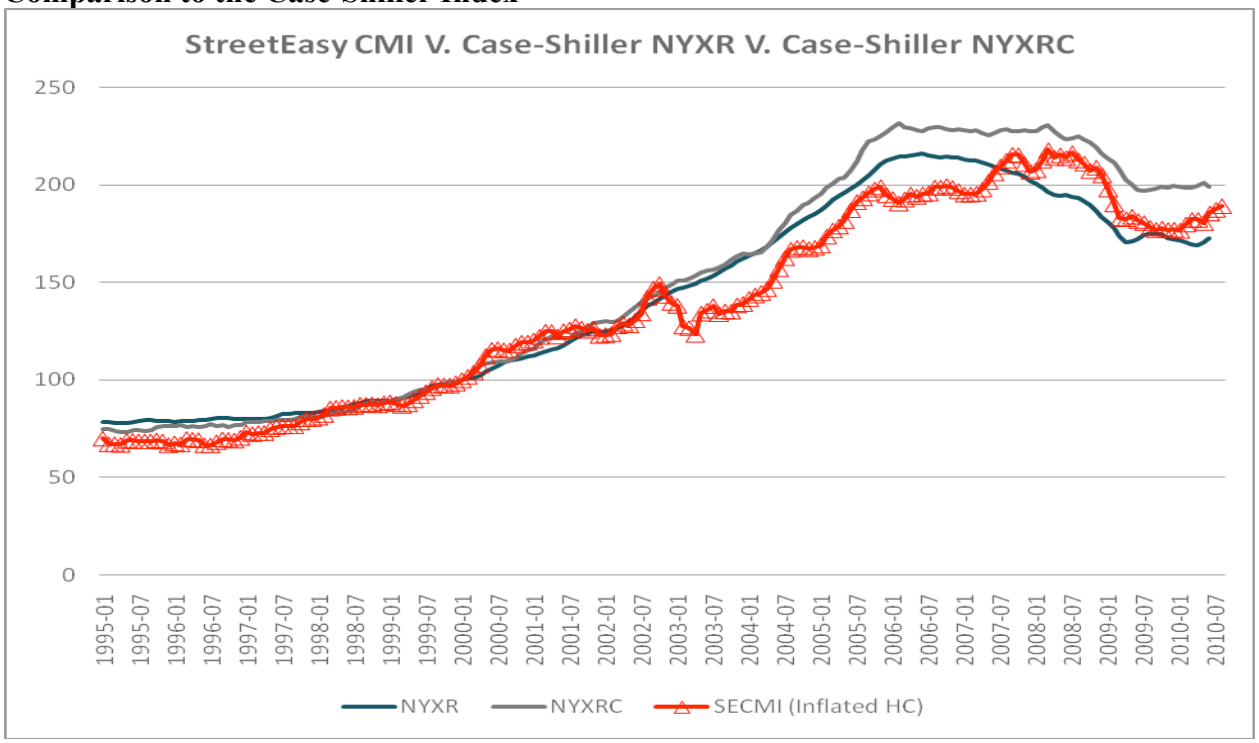
Variations of the StreetEasy CMI

The StreetEasy CMI was manipulated a few additional ways for comparative analysis:

[5] Quigley, John M., 1995, "A Simple Hybrid Model for Estimating Real Estate Price Indexes," Journal of Housing Economics 4, 1-12.

1. Inflation: Pricing trends did not reflect inflation and its influence on nominal price. So we adjusted the model for inflation in order to compare real growth in price v. nominal growth in price by using the following:
    a. Real Price = Nominal Price / Inflation rate (CPI)
2. Given the large data size, we were able to create a major-market index in addition to the overall Manhattan market. However, we were unable to create an index for the smaller, individual neighborhoods, as the sample size would have been too small and not statistically sound. The major markets include:
    a. Downtown – 34th Street and below
    b. Midtown – 34th Street to 59th Street
    c. Upper West Side – West 59th Street to West 110th Street, between Central Park West and the Hudson River
    d. Upper East Side – East 59th Street to East 96th Street between Fifth Avenue and the East River

Comparison to the Case-Shiller Index



In order to compare the StreetEasy CMI to the New York Metro CSI (NYXR) and the New York Metro CSI for Condos (NYXRC), we adjusted our index to have a base year of 2000 like that of Case-Shiller's with a value of 100. Since the CSI is not adjusted for inflation, we removed our inflation adjustment so that we could compare "like-to-like." We found that the StreetEasy CMI showed more price movements than the Case-Shiller indices.

Starting in 2000, the StreetEasy CMI shows that the Manhattan condo market's prices were increasing at a higher rate than both Case-Shiller metro-New York indices, up until September 2001. Additionally, both Case-Shiller indices show that the metro New York market peaked in

early 2006 (NYXR peaked in June 2006 with an index value of 215.83 and the NYXRC peaked in February 2006 with an index value of 231.54). They do not reflect Manhattan's large condo boom seen in 2007 and 2008 where new development condo transactions made up more than one-third of all Manhattan residential closings during that two-year period which caused prices to increase more dramatically. In contrast, the StreetEasy CMI shows that the Manhattan market peaked in March 2008.

The graph clearly illustrates how different the Manhattan condo market is. The NYXR and the NYXRC cover the entire New York metropolitan area, which has been defined to include: Fairfield CT, New Haven CT, Bergen NJ, Essex NJ, Hudson NJ, Hunterdon NJ, Mercer NJ, Middlesex NJ, Monmouth NJ, Morris NJ, Ocean NJ, Passaic NJ, Somerset NJ, Sussex NJ, Union NJ, Warren NJ, Bronx NY, Dutchess NY, Kings NY, Nassau NY, New York NY, Orange NY, Putnam NY, Queens NY, Richmond NY, Rockland NY, Suffolk NY, Westchester NY, Pike PA.[6] With such a large geographic coverage, the CSI does not wholly capture the price movements of a unique market like that of Manhattan.

Further Research

We will continue to explore ways to improve the StreetEasy CMI. In particular, the next two main areas we would like to investigate are:

1. Adjusting the model with more timely data. Perhaps in addition to using recorded sales data from the New York City Department of Finance, we can use sales listings data and analyze time on market and price changes made to listings before they go into contract.

2. Adjusting the model for seasonality. The real estate market in Manhattan is known to be seasonal, with slow summers and winters. Late winter/early spring is traditionally the season for peak contract activity while late spring/early summer is the season for peak closing activity.

Index Update

The index will be updated monthly at the end of each month, as new closings data is acquired. The index values of the previous 24 months will also be revisited and may be updated based on the new data received.

[6] S&P/Case-Shiller Metro Area Home Price Indices Index Methodology, p8. http://www2.standardandpoors.com/spf/pdf/index/SPCS_MetroArea_HomePrices_Methodology.pdf

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 Stratistical Association* 58, 933-942.

Case, E. Karl, and Robert J. Shiller, 1987, "Prices of Single Family Homes Since 1970: New Indexes for Four Cities," *New England Economic Review*, Sept./Oct. 45-56.

Chau, K. W, 2006, "Index Construction Method for The University of Hong Kong All Residential Price Index (HKAPI) and its sub-indices(HKU-HRPI, HKU-KRPI, HKU-NRPI)", Versitech Limited, *The University of Hong Kong, Hong Kong.*

Chihiro Shimizu, Kiyohiko G. *Nishimura, and Tsutomu Watanabe, 2009, House Prices in Tokyo: A comparison of Repeat-Sales and Hedonic Measures*, Understanding Inflation Dynamics of the Japanese Economy, funded by JSPS Grant-in-Aid for Creative Scientific Research, Hitotsubashi University.

Clapham, Eric, Englund, Peter, Quigley, John M., and Redfearn, Christian, 2004, *Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives, University of Cailfornia at Berkeley.*

Goodman, Allen C. and Thibodeau, Thomas G., 1998, "Dwelling Age Heteroskedasticity in Repeat Sales House Price Equations," Real Estate Economics, V26 1:pp. 151-171.

Messner, Chris and Satchell, Stephen, 2010, A comparison of the Case-Schiller house price index methodology with the acadametrics house price index methodology, Acadametrics, UK.

Quigley, John M., 1995, "A Simple Hybrid Model for Estimating Real Estate Price Indexes," Journal of Housing Economics 4, 1-12.

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

Yiu, C. Y. and Tam, C. S., 2004, *The estimation of housing price gradients: A comparison of different approaches applied in Hong Kong*.