House Price Index Methodology

Article · January 2011

3 authors, including:

Chaitra H. Nagaraja
Fordham University
12 PUBLICATIONS   **112** CITATIONS

SEE PROFILE

Susan M. Wachter
University of Pennsylvania
147 PUBLICATIONS   **4,879** CITATIONS

SEE PROFILE

House Price Index Methodology

Chaitra H. Nagaraja,[*] Lawrence D. Brown,[†] Susan M. Wachter[‡] [§]

June 1, 2010

Abstract

This paper examines house price index methodology and explores what makes makes an index both practical and representative. Two approaches are investigated: predictive ability (quantitative) and index structure (qualitative). Five indices are analyzed, four of which are repeat sales indices in the traditional sense and an autoregressive index which makes use of the repeat sales idea but includes single sales as well. The autoregressive index has the best predictive performance.

Contents

[*]Statistical Research Division, U.S. Census Bureau, chaitra.nagaraja@census.gov

[†]Department of Statistics, The Wharton School, University of Pennsylvania, lbrown@wharton.upenn.edu

[‡]Department of Real Estate, The Wharton School, University of Pennsylvania, wachter@wharton.upenn.edu

1 Introduction

Over the past few decades, a number of indices have emerged as more people have looked to the housing market for investment opportunities. In addition, with the current market collapse, housing indicators have become increasingly important in the quest for understanding how such markets operate. The question that still remains is: How do we know how good an index is? Clapp, et al write that "...a residential price index should represent the price change experienced by a typical house within the geographical area covered by the index [6, p. 270]." The most simple index, is that of a median index such as one published by the National Association of Realtors. However, this index is subject to criticism.

The types of homes sold at different times may vary. Therefore changes in the reported index between times (seasons or even across years) may be due to the different composition of homes sold rather than reflecting real changes in the housing market. While it is possible to apply smoothing procedures to mitigate this issue; however, there are a few more criticisms of median indexes. New homes tend to be more expensive than equivalent older homes and it has been suggested that including these into the median index means that the price index will be biased upwards. Furthermore, all sales are treated as if they were single sales in a median price index. However, many homes sell multiple times in the time period–this information on repeat sales is ignored with a mean or median price index. Note that a mean house index is also subject to the same composition problem as the median index.

Bailey, Muth, and Nourse (1963), introduced the landmark concept of repeat sales analysis. Assuming a house has no changes made to it, to assess how prices change over time, one need only to look at the difference in sale prices of the same house. This approach solves the issue of varying composition which mean and median indices suffer from. Subsequent researchers have expanded upon this idea by incorporating various additional features, in an effort to improve index estimates. The most significant, and widely used, development was by Case and Shiller (1987, 1989) who argued that gap times between sales have an effect on sale price differences. The Case and Shiller method is used to compute the Conventional Mortagage Home Price Index released quarterly by Freddie Mac and Fannie Mae. These set of indices cover numerous US cities and regions [15].

There has been much criticism of repeat sales methods, the main issue being that repeat sales indices omit homes that sell only once from analysis which, most importantly, includes new home sales. As a result, the indices are computed from a relatively small subset of all home sales. Consequently, the indices may be unrepresentative of the housing market as a whole especially since new homes tend to be more expensive than older ones. Despite this concern, such procedures have been wholeheartedly adopted by the real estate sector. A number of agencies, including Standard and Poor's and the Office of Federal Housing Enterprise (OFHEO), release indices based on the Case-Shiller method.

We examine five repeat sales indices here: the Bailey, Muth, and Nourse index (hereby referred to as the BMN), the Case and Shiller method (C-S), the Home Price Index produced by the Office of Federal Housing Enterprise Oversight (OFHEO), the S&P/Case-Shiller Home Price Index (S&P/C-S), and an alternative repeat sales index, the autoregressive index (AR

index) introduced by Nagaraja, Brown, and Zhao (2010). This index combines information from both repeat and single sales, with the former being given a higher weight; moreover, advantageous properties follow naturally from the statistical model. For comparison purposes, we also include the median price index as a baseline index.

The goal of this paper is to investigate which useful characteristics each house price index has. As each index employs different techniques, it is important to know whether these variations have an effect on the resulting indices and predictions. We will evaluate these indices using a two-pronged approach: (a) analyzing the components of each index along with the statistical structure and (b) comparing estimates of individual house prices from each index [1].

The second approach requires some justification. Generally, to determine how well a model works for its prescribed purpose, we check with the "truth" either through real data or through simulation. Neither of those techniques can be used here. We do not know the true index value so we cannot compare our estimated index with this; on the other hand, each of these indices assumes a different model for house prices so we cannot use simulation to compare indices. A third option is to examine predictions of the individual house prices as a way of determining the efficacy of the index. The claim here is that methods which produce better predictions, are better models, and thus have more accurate indices. The other approach are more theoretical; comparing predictions is a quantitative method of analysis.

We start in Sec. 2 with a brief description of each index and in Sec. 3, a review of criticisms of repeat sales methods which have been raised in the literature. A technical analysis of the statistical properties of each index can be found in the appendix. The second part of this paper focuses on the practical differences among the five methods. The data used are from home sales from twenty US cities during the period of July 1985 through Septembter 2004. We use these data in Sec. 4 to compare the indices produced from each method and prediction of house prices using each method. We conclude in Sec. 5.

2 Background and Literature Review

The Bailey, Muth, and Nourse method (1963) uses linear regression to compute price index values by utilizing log prices differences between pairs of sales of a house. Essentially the log price difference between a pair of sales is thought to equal the difference in the respective log indices in addition to a homoscedastic error term. Therefore, only houses which have been sold twice are used to calculate the index; the remaining observations are omitted. Homes which are known to have undergone significant improvement or degradation are in principle also excluded from the analysis. This is because for such homes, the previous sale price is

---

[1]In this step, we do not attempt to forecast future individual home prices. None of the methods have this feature. Rather, house prices will be estimated using the fitted model. In essence we are predicting those prices and will use the terms "estimation" and "prediction" interchangeably in this context.

not considered an appropriate surrogate for hedonic information[2].

Case and Shiller (1987, 1989) expand the BMN setup by assuming that the error terms are heteroscedastic. They reason that the length of time between sales should increase the variance of the log price differences between sale pairs. To compute the house price index while accounting for heteroscedasticity, they follow the BMN procedure but add a small twist: when constructing estimates, the observations are weighted depending on the gap time between sales. Sale pairs with larger gap times are given lower weights. There are two independent components to this variance: a fixed component and a variable component which grows as the gap time increases.

The Office of Federal Housing Enterprise Oversight (OFHEO) releases a repeat sales index, the House Price Index (HPI) which is based on the BMN method. Like the Case-Shiller method, a heteroscedastic error term is incorporated but the form of the error term is different. One difference is that the fixed component of the variance is eliminated–this is to avoid problems in the weighting step of the estimation procedure. The second is that the error terms are not independent across multiple sales of the same house. The effects of these alterations will be discussed in more detail in Appendix A.

Standard and Poor's publishes the S&P/Case-Shiller Home Price Index which is based on the arithmetic index proposed by Shiller (1991). This index uses sale prices instead of log prices and is not strictly based on price differences like the indices based on the BMN method. The justification for this change is an easier interpretation of the index in addition to being able to treat houses differently based on the initial sale price [13]. The error structure of this model is nearly identical to the original Case-Shiller method. That is, this model also incorporates a heteroscedastic error term which grows with gap time.

The final index we will examine is an autoregressive index proposed by Nagaraja, Brown, and Zhao (2010). This index is computed using *all* sales; however, as described below, repeat sales are given more influence on the index because more is known about the house when it has sold multiple times. In this conceptual sense, the autoregressive index is a repeat sales index even though it is not based on the BMN methodology. This model is made up of three components: an index, the effect of a home being in a particular ZIP code, and an underlying AR(1) time series which automatically adjusts for the time gap between sales. The ZIP code is included as an additional indicator of its hedonic value. This indicator has some predictive value, although its value is quite weak by comparison with the price in a previous sale, if one has been recorded. Consequently, the estimator that corresponds to this statistical model can be viewed as a weighted average of estimates from a single sale and repeat sales homes, with the repeat sales prices having a dramatically higher weight. As noted, the time series feature of the model guarantees that the weight for repeat sales prices slowly decreases in a natural fashion as the gap time increases. This model incorporates the effect of gap time in two ways. It is included directly because of the underlying AR(1) time series as indicated above and through the variance of the error term which grows as the gap time increases.

---

[2]Hedonic information includes characteristics about a home such as the number of bedrooms and the square footage.

4

The main differences among these five methods are:

1. The BMN model is the only method which assumes the errors are homoescedastic; in the house price setting, this means that the errors do not depend on the gap time between sales.

2. The C-S, S&P/C-S, and OFHEO indices weight observations differently depending on assumptions about the variance of the error terms.

3. The S&P/C-S method models prices instead of differences in log price between sales of the same house and uses instrumental variables in the analysis.

4. The AR method uses an underlying autoregressive time series approach, includes single sales, and location information (ZIP code).

In the following section, we analyze these differences in more detail. For a full technical discussion, see Appendix A.

3 Criticisms of Existing Repeat Sales Methods

While traditional repeat sales methods have proved useful, a number of problems have been highlighted. Perhaps the most obvious issue is that single sales are excluded, thus reducing the sample size significantly. Sample sizes of data used in the actual Case and Shiller (1989) and Meese and Wallace (1997) papers are shown in Table 1. The number of observations which are eliminated is staggering. While data spanning a longer period will result in a higher number of repeat sales, the number of newly built houses also increases. Therefore, the proportion of repeat sales among all house sales does not increase as fast as one might expect.

A second possible disadvantage when single sales are ignored is that in order to have enough repeat sales, you must have a lot of sales to begin with. In other words, repeat sales methods may only apply to large metropolitan areas. However, there is a housing trend for all levels of geography, including local ones. For these smaller areas, there simply may not be enough sales to construct a reasonable repeat sales index. The autoregressive index alleviates this issue by including all sales in the analysis. Therefore, data are not ignored and the index can be applied to any level of geography provided there are enough total sales, not just repeat sales.

Among repeat sales homes, further cuts should be made if the house has significantly improved or deteriorated between sales. This is because "house quality" would not have been controlled in the intervening period. Quite possibly, the Case-Shiller percentages are lower in Table 1 than those for Meece-Wallace because houses that underwent significant renovation or were not "arms-length transactions" (i.e. houses sold cheaply to relatives, etc.) were excluded, eliminating even more data.

A related issue is that in repeat sales models, a home is ignored *until* it is sold for a second time. To see the impact, say that a house is sold first at time t and again at time t'. If the

Table 1: Removing Single Sales

City	No. Obs.	No. Repeat Sales	% Repeat Sales
Case-Shiller (1970-1986)			
Atlanta, GA	221,876	8,945	4.0%
Chicago, IL	397,183	15,530	3.9%
Dallas, TX	211,638	6,669	3.2%
San Francisco/	121,909	8,066	6.6%
Oakland, CA			
Meese-Wallace (1970-1988)			
Freemont, CA	23,408	3,405	14.5%
Oakland, CA	27,606	3,342	12%

initial data set includes sales until time $t^\star$ where $t \leq t^\star < t'$, this particular house would neither be part of the data nor would it be used to compute the index value for time t. Now, if the data is updated to include all sales up to time t', the house is included and is used to compute the index for *both* time t *and* time t'. Thus, indices must be revised retroactively. This can be problematic if indices are to be used in a commercial setting. S&P/Case-Shiller® use a "chain-weighting" procedure to avoid revising the indices [14, p. 26].

Omitting single sales raises the issue whether repeat sales homes are representative of the entire housing market. In any time span, houses can be categorized as follows: new home sales, repeat sales with no changes in the house, repeat sales homes with changes, and houses not sold [3, p. 290]. Repeat sales methods only use data in the second category. If repeat sales homes are far different from other homes, then indices derived from them can only tell us about changes in repeat sales homes, not the entire housing market.

If the set of single sales homes were similar the set of repeat sales homes, then the BMN and succeeding indices would represent the housing market as a whole. The repeat sales method would control for quality of a house along with the issue of varying composition of houses sold over time. However, previous research has shown that this may not be the case.

In an entire sample period, a single sale can refer to a new home or an old home which sells only once in the sample period. There is a hypothesis that a higher proportion of repeat sales are "starter homes." Young families tend to live in these so called "starter homes" and later trade up to larger homes after only a few years [6, p. 271]. Clapp, et al. (1991) test this hypothesis with inconclusive results. Meese and Wallace (1997) test this claim as well by comparing hedonic models using an indicator for repeat sales and incorporating interaction terms between repeat sales and hedonic variables. In their analysis, they found there was a significant difference between repeat and non-repeat sales homes and "...repeat-sales homes that did not change attributes are slightly smaller, and are in worse condition, than the average for single-sale homes...The repeat-sales homes that did have attribute changes...tend to be slightly larger and in worse condition, than the average for single-sale homes [9, 55]."

There is one aspect that varies with all houses: age. Case, et al (1991) claim that because

Table 2: House and Sale Counts

City	No. Sales	No. Houses	Training Pairs	Test
Ann Arbor, MI	68,684	48,522	10,431	9,731
Atlanta, GA	376,082	260,703	59,222	56,127
Chicago, IL	688,468	483,581	105,708	99,179
Columbia, SC	7,034	4,321	1,426	1,287
Columbus, OH	162,716	109,388	27,601	25,727
Kansas City, MO	123,441	90,504	16,705	16,232
Lexington, KY	38,534	26,630	6,075	5,829
Los Angeles, CA	543,071	395,061	75,660	72,350
Madison, WI	50,589	35,635	7,714	7,240
Memphis, TN	55,370	37,352	9,372	8,646
Minneapolis, MN	330,162	240,270	46,206	43,686
Orlando, FL	104,853	72,976	16,147	15,730
Philadelphia, PA	402,935	280,272	63,082	59,581
Phoenix, AZ	180,745	129,993	25,830	24,922
Pittsburgh, PA	104,544	73,871	15,891	14,782
Raleigh, NC	100,180	68,306	16,372	15,502
San Francisco, CA	73,598	59,416	7,111	7,071
Seattle, WA	253,227	182,770	35,971	34,486
Sioux Falls, SD	12,439	8,974	1,781	1,684
Stamford, CT	14,602	11,128	1,774	1,700

age changes over time, repeat sale indices are biased. The time effect is confounded with the age effect. Specifically, the general upward trend of the effect of time is countered by the negative effect of age. As a result, no home will be exactly the same at any two points in time, essentially violating the repeat sales assumption. Palmquist (1979) suggests adding in a depreciation factor to the repeat sales procedure to account for this; however, this factor must be independently computed which adds much complexity to the model [12, p. 337].

4   Comparing Repeat Sales Indices

In this section, we will explore the differences among the indices in an applied setting: in terms of index construction and predictive performance.

The data are comprised of single family home sales from the twenty US metropolitan areas listed in Table 2. These sales occurred between July 1985 and September 2004 and were for homes which qualified for conventional mortgages. The total number of sales and number of unique houses in the data are provided in Table 2 as well. For each sale, the following information is available: address, month and year of sale, and price. We divide the sample period into three month intervals so there are enough sales at each period to compute a

stable index. In total, there are 77 periods, or quarters.

None of our analyses involve deleting repeat sales of significantly renovated homes as is recommended by Case and Shiller. This is because we do not have indicators for such events in our data set. Dropping such homes from the analysis could be expected to improve the performance of *all* repeat sales indices, including the AR index.

We divide the data described above into training and test sets. Each model is applied to the training data. The test set contains the final sale of homes which sell three or more times in the sample period. For homes that sell only twice in the sample period, the final sale is added to the test set with probability 0.5. Roughly 15% of the observations are in the test set. The last two columns of Table 2 provide the size of the training and test sets. The test data area used for prediction.

Prediction Results

Each method is fit using the training data set and the sale price for each house in the test set is predicted. To compare performance across methods, we use the root mean square error (RMSE) defined as: $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(P_i - \hat{P}_i \right)^2}$ where P_i is the sale price of house i, $\hat{P}_i$ is the predicted price, and n is the number of observations in the test set.

In Table 3, we provide the test RMSE results for each method. There is no RMSE value for the S&P/Case-Shiller method for Kansas City, MO. At the second step of the procedure, some of the computed weights were negative preventing the final index values from being computed. Recall, that this is the type of problem that the OFHEO method tries to avoid.

The main feature to note here is that the four traditional repeat sales methods have RMSE values which are quite similar to each other. Hence, the "improvements" made to the BMN model by the other methods seem to result in only minor changes to the RMSEs. However, the autoregressive model clearly has the lowest RMSE values of all.

The RMSE value provides us with information about the following question: On average how well does this model do when applied at the micro level? Prediction of individual house prices is not the end goal for this test. Rather, it is a way to measure the effectiveness of the method in capturing market trends–that is, when constructing an index. Being able to measure market trends more accurately should result in better overall predictions. Using this metric, the autoregressive index performs best.

Index Comparison

Despite the differences in methodology, the indices[3] track each other exceptionally well, even if the median price index is included. The correlations are given in Table 4 for Minneapolis, MN (for clarity, we have removed the bottom half of the table). The high correlations indicate that the general trends match across indices; however, if we plot the indices for three sample cities Atlanta, GA, Minneapolis, MN, and Pittsburgh, PA as in Figs. 1-3, we

---

[3]The indices were computed using the estimates from the training set data.

Table 3: Test Set RMSE (in dollars)

Metropolitan Area	BMN	C-S	S&P/C-S	OFHEO	AR
Ann Arbor, MI	53,709	53,914	52,718	54,024	**41,401**
Atlanta, GA	35,456	35,494	35,482	35,503	**30,914**
Chicago, IL	42,923	42,960	42,865	42,976	**36,004**
Columbia, SC	42,207	42,263	42,301	42,290	**35,881**
Columbus, OH	30,550	30,543	30,208	30,545	**27,353**
Kansas City, MO	27,682	27,724	–	27,730	**24,179**
Lexington, KY	21,748	21,740	21,731	21,741	**21,132**
Los Angeles, CA	41,918	41,949	41,951	41,959	**37,438**
Madison, WI	30,979	30,942	30,640	30,950	**28,035**
Memphis, TN	25,311	25,306	25,267	25,311	**24,588**
Minneapolis, MN	35,402	35,538	34,787	35,565	**31,900**
Orlando, FL	30,187	30,215	30,158	30,228	**28,449**
Philadelphia, PA	35,308	35,333	35,350	35,338	**33,246**
Phoenix, AZ	29,295	29,334	29,350	293,56	**28,247**
Pittsburgh, PA	30,732	30,812	30,135	30,858	**26,406**
Raleigh, NC	26,873	26,856	26,775	26,855	**25,839**
San Francisco, CA	50,513	50,573	50,249	50,499	**49,927**
Seattle, WA	43,533	43,606	43,486	43,631	**38,469**
Sioux Falls, SD	21,527	21,576	21,577	21,525	**20,160**
Stamford, CT	67,661	67,668	68,132	67,579	**57,722**

Table 4: Correlation Among Indices for Minneapolis, MN

	BMN	C-S	S&P/C-S	OFHEO	AR	Median
BMN	1.0000000	0.9999690	0.9999458	0.9999556	0.9995487	0.9933425
C-S	–	1.0000000	0.9999734	0.9999986	0.9995790	0.9931094
S&P/C-S	–	–	1.0000000	0.9999715	0.9994943	0.9928684
OFHEO	–	–	–	1.0000000	0.9995824	0.9930688
AR	–	–	–	–	1.0000000	0.9957070
Median	–	–	–	–	–	1.0000000

can see that the actual value of each index differs. The plot on the left graphs the index produced from each method; the plot on the right is the average index at time t subtracted from the average index level at time t; from this plot, differences among the indices can be more easily detected.

The BMN, C-S, and OFHEO indices are nearly all the same; this is not surprising as the methodology used is nearly identical. The median, S&P/C-S, and AR, indices tend to differ from the others; nevertheless, none of these indices is consistently higher or lower than the BMN, C-S and OFHEO indices. We do note that the AR index is generally between the median index and the traditional repeat sales indices. This is most likely because the median index treats all observations as single sales and the repeat sales indices only include repeat sales homes; the AR index, on the other hand, includes both repeat and single sales. The repeat sales information, however, impacts the index more than the single sales [11].

This lack of pattern could reflect varying growth rates across the twenty metropolitan areas. In Fig. 4, the percentage of repeat sales homes, a home which has been sold in a previous quarter within the sample period, at each quarter is plotted for a selection of cities. As expected, the percentage of repeat sales homes increases as we move through time. In the long run, nearly all homes which appear as single sales will be sales of new homes, not homes which have simply not sold in the sample period. The rates of increase differ widely across cities. After nearly 20 years, the percentage of repeat sales homes is the lowest for San Francisco, CA at 41% and the highest for Columbia, SC at nearly 86%. However, in our data, we cannot distinguish between repeat sales homes and those which are old but have not sold in the sample period; therefore more data are required to determine how differences in growth rates across cities affect the indices.

5   Conclusion

The five indices, BMN, Case-Shiller, OFHEO, S&P/Case-Shiller, and the autoregressive index are all based upon the repeat sales idea: that the previous sale price contains all of the information about a house. The way this information is used is what differentiates these indices. The BMN index assumes that the gap time between sales does not affect the value of this previous sale price. The Case-Shiller, OFHEO, and S&P/Case-Shiller
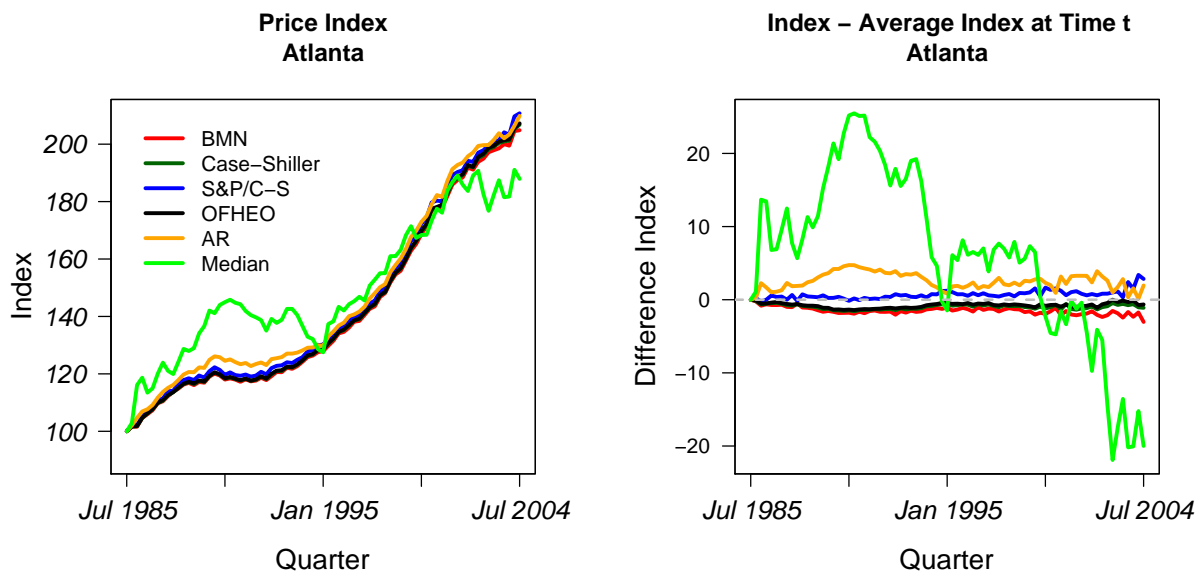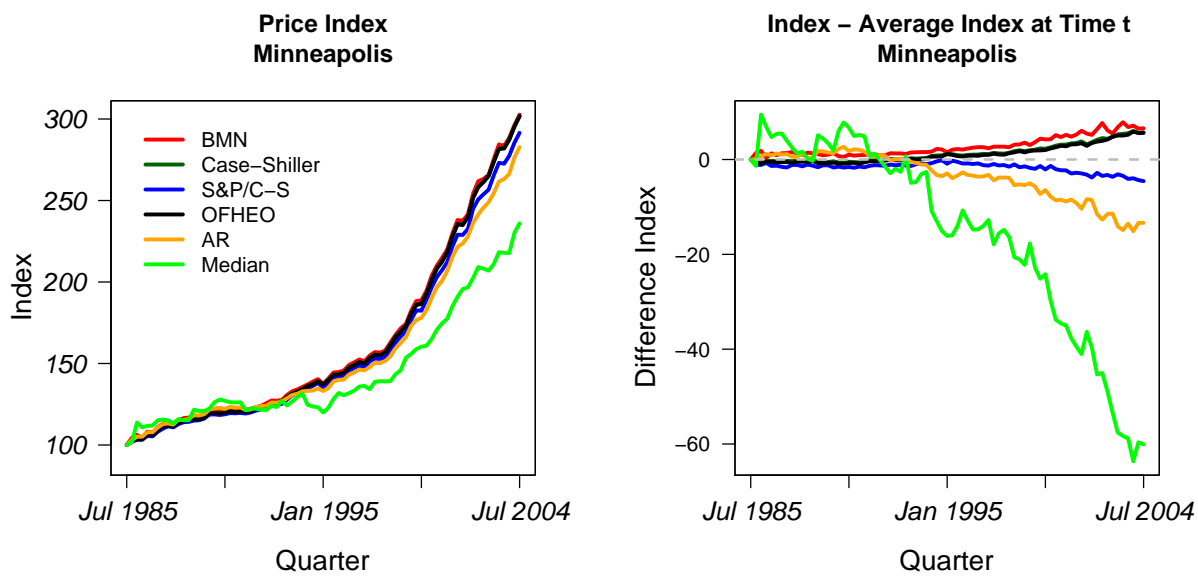
Figure 1: Indices for Atlanta, GA

**Price Index
Atlanta**

Legend: BMN, Case–Shiller, S&P/C–S, OFHEO, AR, Median

**Index − Average Index at Time t
Atlanta**

Figure 2: Indices for Minneapolis, MN

**Price Index
Minneapolis**

Legend: BMN, Case–Shiller, S&P/C–S, OFHEO, AR, Median

**Index − Average Index at Time t
Minneapolis**

11

Figure 3: Indices for Pittsburgh, PA

**Price Index
Pittsburgh**

**Index – Average Index at Time t
Pittsburgh**
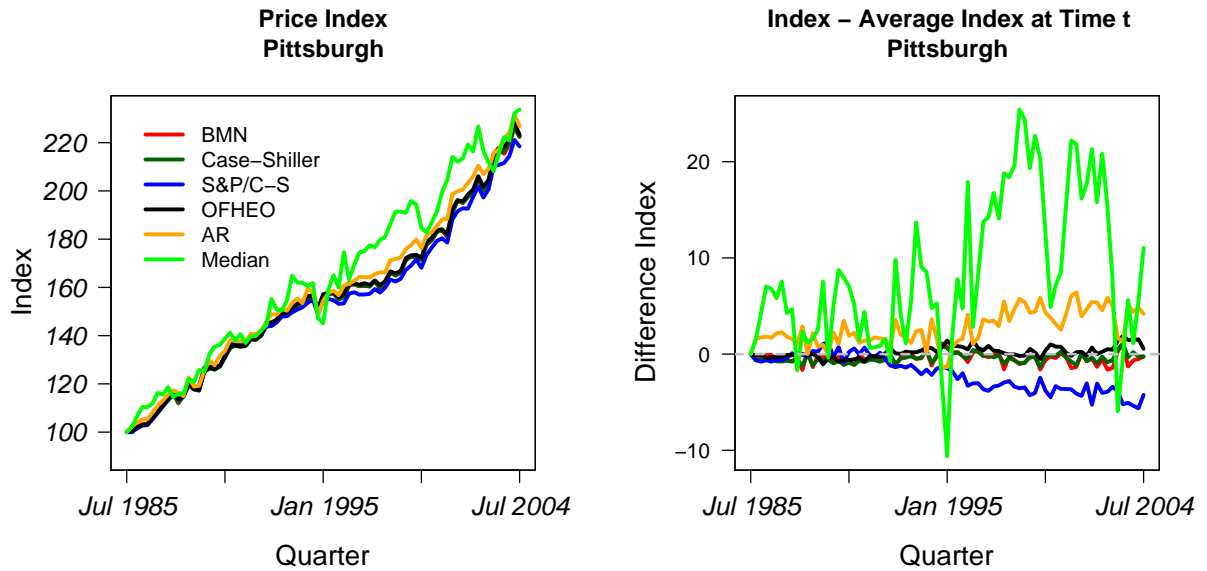
Figure 4: Percentage of Repeat Sales by Quarter for a Selection of Cities

Percent Repeat Sales Homes by Quarter
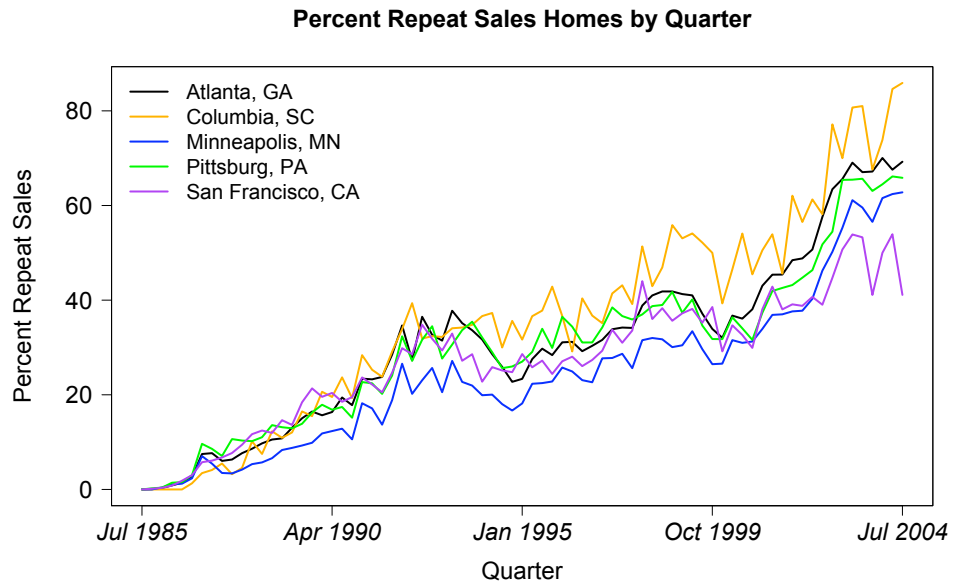
indices assume that the gap time affects the random variation component of the regression model. The autoregressive method takes this one step farther. Not only does the gap time increase the error variance, it also decreases the covariance between the sale prices. The higher the gap time, the lower the covariance between sale pairs. The Case-Shiller based models also decrease the covariance but only by increasing the error variance. Furthermore, the autoregressive method assumes that the previous price *combined* with the ZIP code (geographical information) provide all of the information about a house.

Each of the traditional repeat sales indices attempt to correct problems from preceding indices. The resulting indices do differ substantially as seen in the previous section. However, when comparing predictive power, none of these indices consistently outperform the others leading one to feel that these improvements do not provide any substantive benefits. Furthermore, the statistical implications of these improvements are often disregarded. As a result, standard deviations and confidence intervals constructed from the results are not accurate calculations of the variation in index estimates. These criticisms apply to the Case-Shiller, OFHEO, and S&P/Case-Shiller indices.

When one has to choose among these indices, the question is: what makes a good house price index? If usability is key, all of these indices are good–all are easy to implement and update and do not require much information about a house. If statistical properties are important, the BMN and autoregressive indices are best. A third measure is how well the index represents trends in the overall market. Previous research has shown that repeat sales homes are fundamentally different from single sales; in light of this work, it is difficult to argue that such indices can truly represent the housing market. While the median index and autoregressive indices exclude houses that do not sell, they do include single sales which can make up a large proportion of total sales. Therefore, in this regard, these two indices are more representative of the housing market. Furthermore, the autoregressive index makes better use of the data by taking advantage of the additional information contained in repeat sales. However, none of these standards indicates whether an index is truly measuring what is supposed to. We feel the best yardstick in this regard is predictive ability. In this case, the autoregressive index is the clear winner. In fact, the autoregressive model seems to best embody what an index should represent and what housing indices are actually used for.

Acknowledgements

A   Structure

In this section, we provide a more technical description of each model discussing the statistical properties of each method.

A.1 The Bailey, Muth, and Nourse Model

Let there be $T+1$ time periods where sale can occur from $0, 1, \ldots, T$ and t be the subscript for time. Using the BMN notation, for a pair of sales of a given house i, prices and indices are related by the following expression:

$$\frac{P_{it'}}{P_{it}} = \frac{B_{t'}}{B_t} U_{itt'} \tag{1}$$

where P_{it} is the sale price of the ith house at the tth time period. Let t be the time at the first sale and t' the time at the second $(t' > t)$ and let B_t denote the general house price index at time t. Finally, $U_{itt'}$ is the error term and is assumed to have a log normal distribution: $\log U_{itt'} \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$ where iid denotes independent and identically distributed [1, p. 934] . The model is then fit on the logarithmic scale:

$$p_{it'} - p_{it} = b_{t'} - b_t + u_{itt'} \tag{2}$$

where p, b, and u are simply the logarithmic versions of the terms in (1).

Essentially, the expected difference in log prices for two sales of a house is thought to equal the difference in the corresponding log indices. The varying gap times between sales is considered irrelevant and as a result, the error terms are asumed to be homoscedastic. This issue is addressed in various ways in subsequent repeat sales indices.

A.2 The Case and Shiller Method

Like the BMN model, the Case and Shiller (C-S) setting is a model for differences in log prices. Thus, we can simply adjust the model in (2) to incorporate an alternate error structure:

$$p_{it'} - p_{it} = b_{t'} - b_t + H_{i,t'} - H_{i,t} + u_{it'} - u_{it}$$
$$= b_{t'} - b_t + \sum_{j=t+1}^{t'} v_{ij} + u_{it'} - u_{it} \tag{3}$$

where $t' > t$ and $H_{i,t}$ is a Gaussian random walk. Therefore, $H_{i,t'} - H_{i,t}$ is simply the sum of the intervening steps of the random walk: $\sum_{j=t+1}^{t'} v_{ij}$. Case and Shiller assume that $u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$, $v_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2)$, and that u_{it} and v_{it} are independent of each other for all time periods and houses.

Case and Shiller use weighted least squares to fit the model to account for both sources of variation. Their three-step procedure is described below.

1. Fit the BMN model in (2).

2. Compute the residuals of the regression in (2), and denote these as $\hat{\varepsilon}_{itt'}$. These residuals are an estimate of: $u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}$. The expectation of $\varepsilon_{itt'}$ is $E\left[u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}\right] =$

0 and the variance is $Var\left[u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}\right] = 2\sigma_u^2 + (t' - t)\sigma_v^2$ because the errors are independent of each other. The square of the residuals is an unbiased estimate of this variance. To compute the weights for each observation, the squared residuals from Step 1 are regressed against the gap time. That is,

$$\hat{\varepsilon}_{itt'}^2 = \underbrace{\beta_0}_{2\sigma_u^2} + \underbrace{\beta_1}_{\sigma_v^2}(t' - t) + \eta_{itt'} \tag{4}$$

where $E[\eta_{itt'}] = 0$. The reciprocal of the square root of the fitted values from the above regression are the weights denoted by $\mathbf{W}^{-1}$.

3. To obtain the final index; run a weighted least squares regression. Basically, step 1 is repeated while incorporating the weights. This matrix $\mathbf{W}^{-1}$ reduces the effect of sale pairs with large gap times on the index estimates.

Both the C-S and S&P/C-S indices are computed using generalized least squares (GLS). The standard GLS procedure would be to define the weight matrix $\mathbf{W}^{-1}$ as a matrix of estimated variances. Such weights are generally used so that the best linear unbiased estimates (BLUE) of the regression coefficients are obtained. However, in the C-S models the estimated *standard deviations* are used instead. Consequently, the resulting index estimates are unbiased but do not have the lowest possible variance. This is undesirable especially if the regression estimates are to be used for prediction and prediction intervals are to be constructed.

Another curious feature of the error process in these indices when there are more than two sales of a house. A curious feature of the C-S model arises when examining the case of multiple sales. For instance, say a house is sold thrice: one at time 0, a second at time h, and a third at time $h + g$. Recall that the variance of the difference of a pair of sales is given by $2\sigma_u^2 + (t' - t)\sigma_v^2$ where t and t' are the times when the sales occurred. Say, by chance, we do not know about the second sale. Then, the variance of the difference of the first and third sale should be $2\sigma_u^2 + (g + h)\sigma_v^2$. Ideally, the fact that there was a second sale which was missing from the data should not be informative; that is, the variance of the estimates should not change with this knowledge. However, this is not the solution if derived from the regression equations. Rather, knowing there is a second sale at time h is informative. To see why this is true, we start by writing the regression equations for both pairs of sales:

$$p_{i\,h} - p_{i\,0} = \beta_h - \beta_0 + \varepsilon_{i\,0\,h} \tag{5}$$
$$p_{i\,h+g} - p_{i\,h} = \beta_{h+g} - \beta_h + \varepsilon_{i\,h\,h+g} \tag{6}$$

where $\varepsilon_{itt'}$ includes both the random error and the cumulative random walk error. Adding

(5) and (6), we obtain:

$$
\begin{aligned}
p_{i\,h+g} - p_{i\,h} + p_{i\,h} - p_{i\,0} &= \beta_{h+g} - \beta_h + \beta_h - \beta_0 + \varepsilon_{i\,0\,h} - \varepsilon_{i\,h\,h+g} \\
p_{i\,h+g} - p_{i\,0} &= \beta_{h+g} - \beta_0 + \varepsilon_{i\,0\,h} - \varepsilon_{i\,h\,h+g} \\
Var\left[p_{i\,h+g} - p_{i\,0}\right] &= Var\left[\varepsilon_{i\,0\,h}\right] + Var\left[\varepsilon_{i\,h\,h+g}\right] \\
&= 2\sigma_u^2 + h\sigma_v^2 + 2\sigma_u^2 + g\sigma_v^2 \\
&= 4\sigma_u^2 + (g+h)\sigma_v^2 \quad\quad (7)
\end{aligned}
$$

where $Var[\cdot]$ is the variance function. The variance of the first and third sales, given the knowledge of the second, is *larger* than if we had simply omitted it from the data.

A final issue with the C-S methodology is that in the second stage when weights are computed, there is always a chance that for a particular sale pair, the computed weight may in fact be negative. For such cases, the third step cannot be executed at all. Calhoun (1996) outlines a method to circumvent this issue and is described in the next section.

A.3 The OFHEO HPI Method

The OFHEO index is published for each US state, census division, and nationwide. The data are provided by the Federal Home Loan Mortgage Corporation (Freddie Mac) and the Federal National Mortgage Association (Fannie Mae) and contain homes which qualify for a conventional mortgage. This criterion excludes some high-end homes and homes bought at subprime rates.

The OFHEO index method follows the Case-Shiller method; however, a few adjustments are made. In the second stage of the method when the weight matrix is computed, it is possible for some of the weights to be negative. Recall, the regression in (4) is fitted to calculate the weights. The intercept of this regression could be negative and large enough to offset the second term, resulting in a negative weight. In such situations, the third step of the C-S method cannot be computed as all weights must be non-negative. To eliminate this issue, instead of writing the model as in (3), the white noise component, u_{it} is replaced by u_i. That is, there is only one error term for each house, not for each house and sale combination [2, p. 9]. The resulting model is:

$$
\begin{aligned}
p_{it'} - p_{it} &= b_{t'} - b_t + H_{i,t'} - H_{i,t} + u_i - u_i \\
&= b_{t'} - b_t + \sum_{j=t+1}^{t'} v_{ij} \quad\quad (8)
\end{aligned}
$$

Essentially, u_i is treated as a fixed effect in the linear regression. As the C-S index looks at *differences* between prices, u_i drops out of the fitted model.

In the Case-Shiller procedure, H_{it} is a Gaussian random walk. Therefore, each step of the random walk is assumed to be independent of the previous step. This is not the case for the OFHEO index; the steps are assumed to be dependent. This means that the errors in

the regression when fitting the model are not independent which is a violation of a standard regression assumption.

For two sales of a house at time t and t', recall:

$$H_{it'} - H_{it} = \sum_{j=t+1}^{t'} v_{ij}$$

where v_{ij} are the random walk steps and $E\left[v_{ij}\right] = 0$ where $E[\cdot]$ is the expectation function. If we compute $Var\left[H_{it'} - H_{it}\right]$,

$$
\begin{aligned}
Var\left[H_{it'} - H_{it}\right] &= Var\left[\sum_{j=t+1}^{t'} v_{ij}\right] \\
&= E\left[\left(\sum_{j=t+1}^{t'} v_{ij}\right)^2\right] - E\left[\sum_{j=t+1}^{t'} v_{i,j}\right]^2 \\
&= \sum_{j=t+1}^{t'} E\left[v_{ij}^2\right] + \sum_{j=t+1}^{t'} \sum_{\substack{j'=t+1 \\ j \neq j'}}^{t'} E\left[v_{ij}v_{ij'}\right] \\
&= (t'-t)E\left[v_{ij}^2\right] + (t'-t)((t'-t)-1)E\left[v_{ij}v_{ij'}\right] \\
&= (t'-t)\left(E\left[v_{ij}^2\right] - E\left[v_{ij}v_{ij'}\right]\right) + (t'-t)^2 E\left[v_{ij}v_{ij'}\right] \quad (9)
\end{aligned}
$$

where $j \neq j'$. In the Case-Shiller model, $E\left[v_{ij}v_{ij'}\right] = 0$ for all $j \neq j'$ leaving $Var\left[H_{it'} - H_{it}\right] = \sigma_v^2(t'-t)$. However, this assumption is not made when computing the OFHEO HPI index. Consequently, at the second stage of the fitting procedure, the squared residuals are instead regressed against the gap and squared gap times [2, p. 10]:

$$\hat{\varepsilon}_i^2 = \alpha_0 (t'-t) + \alpha_1 (t'-t)^2 + \eta_i \quad (10)$$

where $\eta_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right)$. Note that if $E\left[v_{it}v_{it'} \neq 0\right]$, H_{it} of (8) is no longer a random walk. This is because the v_{it} values are independent in a random walk so $E[v_{it}v_{it'}] = 0$. This causes a problem statistically when there are more than two sales of a house. For example, say a house sells at times $t_0, t_1, t_2, \ldots$. The log prices of the house $p_{it_0}, p_{it_1}, p_{it_2}, \ldots$ can be thought of as a time series. As repeat sales methods look at sale pairs, this would mean the first and second sale would form a sale pair and the second and third sale would form a sale pair. Therefore, these two sale pairs would have the form:

$$
\begin{aligned}
p_{i1} - p_{i0} &= b_1 - b_0 + v_{i1} \\
p_{i2} - p_{i1} &= b_2 - b_1 + v_{i2}
\end{aligned}
$$

If H_{it} was a random walk, this is not an issue as $E[v_{i1}v_{i2}] = 0$ and the errors v_{it} are not correlated. However, if $E[v_{i1}v_{i2}] \neq 0$, then this regression would have correlated observations,

specifically, autocorrelated errors. Independence of the observations, in this case sale pairs, is a vital regression assumption which is violated.

One consequence of this violation is that while the estimates of the index values are unbiased, they no longer have the minimum variance property. (Recall, they already do not have this property because the weighted least square step uses the *square root* of the estimated variance values instead of the square as weights.) Furthermore, the estimate of the standard errors and any confidence intervals may not be valid.

Calhoun (1996) suggests an adjustment factor that can be applied to the computed index so that the repeat sales index, a geometric index, behaves more like an arithmetic index. Arithmetic indices have the advantage of having an easier cross-sectional interpretation. Goetzmann (1992) first proposed this modification and Stephens, et al (1995) and Calhoun (1996) show how to implement it when applying the OFHEO index to a particular house.

Jensen's inequality states that for a concave function, such as $\log x$, $E\left[\log x\right] \leq \log E\left[x\right]$ where $E[\cdot]$ is the expectation function. If we substitute β_t for x, $E\left[\log x\right]$ is what we compute using the OFHEO index methodology whereas $\log E\left[x\right]$ is what is desired to obtain $E\left[x\right]$. Goetzmann proposed an adjustment to approximate $E\left[x\right]$ from $E\left[\log x\right]$ for the repeat sales case. A common approximation is $E\left[x\right] \approx exp\left\{E\left[\log x\right] + \frac{Var[\log x]}{2}\right\}$. Goetzmann suggests that $E\left[\log x\right]$ is estimated by the log price index $\hat{\beta}_t$ and $Var[\log x]$ is estimated by $\hat{\sigma}_v^2$, the variance associated by the gap time in the original Case-Shiller model [7, p. 9,12].

The variance of the error term for the OFHEO index is quite different than for the Case-Shiller index, Calhoun proposes using (10) in place for $Var[\log x]$ [2, p. 11]. To apply this to individual houses, use $\hat{P}_{t'} = P_t \exp\left\{\hat{\beta}_{t'} + \frac{\alpha_0(t'-t)+\alpha_1(t'-t)^2}{2}\right\}$ [15, p. 393]. However, following this procedure seems to give poor predictions. Therefore, in our data analysis, we skip this adjustment when computing predictions using the OFHEO methodology.

A.4   The S&P/Case-Shiller Method

A variation of the interval and value-weighted arithmetic repeat sales estimator proposed by Shiller (1991) is currently used to construct the S&P/Case-Shiller Home Price Index released by Standard and Poor's [14]. The S&P/C-S index is published for 20 Metropolitan Statistical Areas (MSA) and nationally. For this commercial index, indices are computed using a rolling three-month window. That is, a house sale in October is used for computing the index for October, November, and December. This is done by listing the house three times, for October through December, and weighting each "expanded observation" by 1/3 [14, p. 26].

The construction of an index on log price differences makes the indices considered thus far geometric indices. For a better cross-sectional interpretation, an arithmetic index would be favored argues Goetzmann [7, p. 9]. Goetzmann proposes an adjustment that can be applied to the traditional repeat sales indices which approximate an arithmetic index (mentioned in App. A.3). Shiller, however, proposes an entirely new model which we describe next.

As before, we have $T + 1$ time periods from 0, 1, ..., T. For sale pair i, we can write this

model as:

$$
\begin{aligned}
P_{i0} &= \beta_{t'} P_{it'} + U_{i0t'} & \text{first sale at time 0,} \\
0 &= \beta_{t'} P_{it'} - \beta_t P_{it} + U_{itt'} & \text{first sale at time } t > 0
\end{aligned}
\tag{11}
$$

where P_{it} is the sale price of house i at time t, β_t is the inverse of the index at time t, and $U_{itt'} \overset{iid}{\sim} \mathcal{N}\left(0, 2\sigma_u^2 + (t' - t)\sigma_v^2\right)$ where σ_v^2 and σ_u^2 are the same variances from the original Case-Shiller index. The index $B_t = \frac{1}{\beta_t}$. As before, the price index at time 0 is set to 1 by convention ($B_0 = 1$).

As this index is formulated as an arithmetic index, it is *value-weighted*. As Shiller (1991) argues, "The weighting may make a difference to the estimated index if price changes in more valuable houses are different from price changes in less valuable houses [13, p. 110]."

Observe that the response vector in (11) contains mostly zeros as the vast majority of sales do not occur in the base time period. However, as sales in the base period are the only sales to not be multiplied by an index since $B_0 = 1$ by construction, one must assume that this is why they are the only prices that appear in the response vector. Moreover, it seems misleading to create a model where *future* sales are used to explain a *preceding* sale. For those sale pairs where the first sale is in the base period, this is exactly what occurs.

The S&P/C-S index looks in some ways nearly like the C-S index but on the price scale. The main difference is that the error term, $U_{itt'}$, is not multiplicative but additive for the S&P/C-S index. However, the random walk error component and additional error terms which result from the multiplicative structure of the original model are carried through. Recall, the variance of the error terms in the Case-Shiller is: $2\sigma_u^2 + (t' - t)\sigma_v^2$ which is the random variation for the *difference* of two sales of a house. However, the S&P/Case-Shiller index is not constructed from differences in prices. Meissner and Satchell (2007) observe this issue as well in their paper comparing this index with the Financial Times House Price Index (a non-repeat sales index) used in the United Kingdom.

This model also assumes that prices do not reflect the true value of the house. That is, there is some measurement error. In least-squares regression, the explanatory variables are assumed to be fixed, not variable. Introducing measurement error into the prices violates this assumption resulting in biased coefficient estimates. To accommodate this issue, instrumental variables (IV) are used during model fitting [16, p. 75-78]. This index is fit using a three step process similar to the original Case-Shiller method.

A.5   The Autoregressive Model

A different method of using repeat sales information is to treat sales of a house as a time series. Repeat sales information is automatically incorporated into a model of this type. In particular, let $p_{i,j,z}$ be the log price of the jth sale of the ith house in zip code z. Let $\mu + \beta_{t(i,j,z)}$ be the log price index for time period, $t(i, j, z)$, and let $\gamma(i, j, z)$ be the gap time, or $t(i, j, z) - t(i, j - 1, z)$, if it is the second or higher sale. Finally, zip code is modeled as a

random effect: $\tau_z \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right)$. Then,

$$
\begin{array}{ll}
p_{i,1,z} = \mu + \beta_{t(i,1,z)} + \tau_z + \varepsilon_{i,1,z} & j = 1 \\
p_{i,j,z} = \mu + \beta_{t(i,j,z)} + \tau_z + \phi^{\gamma(i,j,z)}\left(p_{i,j-1,z} - \mu - \beta_{t(i,j-1,z)} - \tau_z\right) + \varepsilon_{i,j,z} & j > 1
\end{array} \tag{12}
$$

The error distributions are as follows: $\varepsilon_{i,1,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$,

$\varepsilon_{i,j,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j,z)}\right)}{1-\phi^2}\right)$, and all $\varepsilon_{i,j}$ are assumed to be independent. The adjusted log price value, $p_{i,j,z} - \mu - \beta_{t(i,j,z)} - \tau_z$, is a first order autoregressive time series with autoregressive coefficient ϕ. Maximum likelihood estimation is used to fit this model. Details can be found in [11].

This model is in essence a repeat sales method as repeat sales information is included through the autoregressive setup. However, single sales are not ignored in computing the index. Rather, the index is computed roughly as a weighted average of both repeat sales and single sale price information with much more weight on the repeat sales homes since more information is known about a house if it has been sold more than once.

A second feature of this model is the way it handles the gap time between sales. First, sale pairs with longer gap times are assumed to be less correlated than those sale pairs with shorter gap times. The latent autoregressive component $\phi^{\gamma(i,j,z)}$ incorporates this feature directly into the prediction process. Furthermore, the variance of predictions with long gap times is much larger than those with short gap times. This is evident in the construction of the error variance: $\frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j,z)}\right)}{1-\phi^2}$. This value increases as the gap time, or $\gamma(i,j,z)$ lengthens.

References

[1] Bailey, M.J., Muth, R.F., Nourse, H.O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association.* **58** 933-942.

[2] Calhoun, C. (1996). OFHEO house price indices: HPI technical description. http://www.ofheo.gov

[3] Case, B., Pollakowski, H.O., Wachter, S. (1991). On choosing among house price index methodologies. *American Real Estate and Urban Economics Association Journal.* **19** 286-307.

[4] Case, K.E., Shiller, R.J. (1987). Prices of single-family homes since 1970: new indexes for four cities. *New England Economic Review.* **Sept./Oct.** 45-56.

[5] Case, K.E., Shiller, R.J. (1989). The efficiency of the market for single family homes. *The American Economic Review.* **79** 125-137.

[6] Clapp, J.M., Giaccotto, C., Tirtiroglu, D. (1991). Housing price indices based on all transactions compared to repeat subsamples. *American Real Estate and Urban Economics Association Journal.* **19** 270-284.

[7] Goetzmann, W.N. (1992). The accuracy of real estate indices: repeat sales estimators. *Journal of Real Estate Finance and Economics.* **5** 5-53.

[8] Kagarlis, M., Peterson, D., Eberlein, R., Fiddman, T. (2007). The Radar Logic™Daily Index.
http://www.radarlogic.com

[9] Meese, R.A., Wallace, N.E. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *Journal of Real Estate Finance and Economics.* **14** 51-73.

[10] Meissner, C., Satchell, S. (2007). A comparison of the Case-Shiller house price index methodology with the FT house price index methodology.
htp://www.acadametrics.co.uk

[11] Nagaraja, C.N., Brown, L.B., Zhao, L. (2009). An autoregressive approach to house price modeling.

[12] Palmquist, R.B. (1982). Measuring environmental effects on property values without hedonic regression. *Journal of Urban Economics.* **11** 333-347.

[13] Shiller, R.J. (1991). Arithmetic repeat sales price estimators. *Journal of Housing Economics.* **1** 110-126.

[14] S&P/Case-Shiller® Home Price Indices (2007).
http://www.standardandpoors.com

[15] Stephens, W., Li Y., Lekkas, V., Abraham, J., Calhoun, C., Kimmer, T. (1995). Conventional mortgage home price index. *Journal of Housing Research.* **6** 389-418.

[16] Stock, J. (2002). Instrumental variables in statistics and econometrics. Entry in *International Encyclopedia of the Social Sciences.* 7577-7582. Elsevier, Amsterdam.