Repeat Sales House Price Index Methodology

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

Fordham University[†], University of Pennsylvania[‡]

Abstract

We compare four traditional repeat sales indices to a recently developed autoregressive index that makes use of the repeat sales methodology but incorporates single sales and a location effect. Qualitative comparisons on statistical issues including the effect of gap time on sales, use of hedonic information, and treatment of single and repeat sales are addressed. Furthermore, predictive ability is used as a quantitative metric into the analysis using data from US home sales in twenty metropolitan areas. The indices tend to track each other over time; however, the differences are substantial enough to be of interest, and we find that the autoregressive index performs best overall.

Keywords: repeat sales, time series, housing, index evaluation

1 Introduction

Housing is an important part of a nation's economy and house price indices help us understand how such markets operate by tracking changes over time. These indices can be useful for a variety of purposes: as macroeconomic indicators, as input into other indicators, by individuals looking to sell or purchase a home, or for appraising homes.

The goal of this paper is conduct a comparative analysis of the autoregressive index introduced by Nagaraja, Brown, and Zhao (2011) with four traditional repeat sales indices: the Bailey, Muth, and Nourse index, the original 1987 Case and Shiller method, the Home Price Index produced by the Federal Housing Finance Agency, and the S&P/Case-Shiller Home Price Index published by Standard and Poor's. For comparison purposes, we also include the median price index. We do not compare results for hedonic indices or the hybrid index from Case and Quigley (1991) because, other than location information, no hedonic information is available in our data.

We will evaluate these five 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. Using data from home sales for twenty U.S. cities, from 1985 to 2004, we find that the autoregressive model produces the best predictions.

The second approach requires some justification. Generally, to determine how well a model works for its prescribed purpose, we check with the "truth" either using real data or through simulation. We cannot apply either of those techniques here. There is no true index and as each index is constructed under differing data generating processes, simulation is not an effective tool for comparison. A third option is to examine predictions of individual house prices as a way to determine the efficacy of an index. All of the indices studied here can be applied on a microeconomic level. Therefore, individual price prediction is a practical way of evaluating an index. This type of quantitative metric allows for comparisons on an objective, measurable scale. We assert in this paper that methods which produce better predictions involve better models and thus lead to more accurate indices. Therefore, prediction, combined with qualitative comparisons, provides a more complete analysis of a housing index.

We begin in Section 2 with a literature review and describe the data in Section 3. In Section 4, we define each of the models and in Section 5 we make qualitative comparisons. We use the data in Section 6 to compare the indices and predictions produced from each method. We conclude in Section 7.

2 Background

A major hurdle in constructing house price indices is that homes are heterogenous goods. Furthermore, the market composition of homes sold changes throughout the year causing even more difficulties. One way to control for differences in the quality of housing stock over time is to use a hedonic index. Characteristics of a house such as floor area or location are considered hedonic variables. Such indices, such as the one proposed in Yeats (1965) and Noland (1979), are constructed by regressing the hedonic characteristics against sale prices. Pure hedonic models have been largely abandoned in favor of alternative methods mostly due to problems with the availability, accuracy, and stability of relevant variables and the difficulty in describing the model form. Other proposed methods, such as repeat sales or spatial models, attempt to circumvent such issues by using previous sale price and geography respectively as surrogates for hedonic variables; however, Meese and Wallace (1997) still advocate the use of hedonic models for constructing local indices.

Bailey, Muth, and Nourse (1963) introduced the landmark concept of repeat sales analysis. Using this method, assuming a house undergoes no changes, to assess how prices change over time, one need only look at the difference in sale prices of the same house. This approach appears to solve the issue of varying composition which mean and median indices suffer from and it addresses the problem of hedonic models not capturing all characteristics. 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. A minor modification of the Case and Shiller method is used to compute the Conventional Mortgage Home Price Index released quarterly by Freddie Mac and Fannie Mae. This set of indices covers numerous US cities and regions.

There has been much criticism of repeat sales methods. A key objection is that repeat sales methods exclude homes which sell only once within the time span of data collection. Proponents of these methods claim that such exclusions are necessary to obtain a "constant-quality" index. They argue that as new homes tend to be of higher quality than old homes, the true effect of time is confounded with quality if included (Case and Shiller, 1987). As a result, the indices are computed from a small subset of all home sales. Consequently, they may be unrepresentative of the housing market as a whole. We find that in our data, the sample size is reduced significantly if only repeat sales homes are included: between 33% (Columbia, SC) and 64% (San Francisco, CA) of the data are single sales homes (see Table 1).

This feature is also well recorded in the literature (see Case, Pollakowski, and Wachter (1991) and Meese and Wallace (1997)). Furthermore, Clapp, Giacotto, and Tirtiroglu (1991) find that repeat sales homes are fundamentally different from single sales homes. Therefore, Clapp and Giacotto (1992), claim repeat sales indices may show changes in repeat sales homes only and not the entire housing market (p. 306). Englund, Quigley, and Redfearn (1999) have similar reservations in their analysis of Swedish home sales. Gatzlaff and Haurin (1997) also find that in an analysis of home sales in Dade County, Florida that repeat sales indices, by virtue of being constructed from a subset of all sales, suffer from sample selection bias. Despite these concerns, such procedures have been widely adopted by the real estate sector. A number of agencies, including the Federal Housing Finance Agency (Calhoun 1996) and Standard and Poor's (2009), constructed using Shiller's 1991 paper on arithmetic indices, release indices based on the repeat sales method.

A second issue is that all houses age and therefore it could be argued that no pair of sales are of identical homes, tempering the constant-quality index argument. Case, et al (1991) claim that because age increases over time, repeat sale indices are biased because time effects are confounded with age effects. Specifically, the general upward trend of the effect of time is countered by the negative effect of age. Furthermore, the age effect is not accounted for in repeat sales indices as shown in Palmquist (1979). Palmquist (1982) suggests adding in a depreciation factor to the repeat sales procedure to account for this; however, this factor must be independently computed adding complexity to the model. Goodman and Thibodeau (1996) propose an iterative repeat-sales method which uses age in the model for estimating the effect of gap time. Clapp and Giacotto (1998b), Cannaday, Munneke, and

Yang (2005), and Chau, Wong, and Yiu (2005) propose models where age is included in such a way as to avoid collinearity with the index component of repeat sales methods. As year of construction is not available in our data, we are unable to implement these methods for comparison purposes for any of the five indices.

Case and Quigley (1991) propose a hybrid index which incorporates hedonic variables so that homes excluded in traditional repeat sales indices are not omitted. Clapp and Giacotto (1998a) construct a repeat sales method which incorporates assessed values of homes as a hedonic variable in addition to actual sales. Gatzlaff and Ling (1994) find that the assessed value method and repeat sales methods produce similar results. Knight, Dombrow, and Sirmans (1995) propose a hybrid index which permit the hedonic coefficients to vary over time. However, the data requirements for any of these hybrid indices may make them impractical to implement on a broad scale.

The autoregressive index proposed by Nagaraja, Brown, and Zhao (2011) incorporates additional data, specifically single sales and ZIP code and, in this way, is similar to hybrid indices. Where it differs is that only sale price and address are required to deploy this index. The advantage of this, like all repeat sales price indices, is that the autoregressive index can be used across broad geographies. House price indices are estimated for the US as a whole, by metropolitan area, as well as for sub-areas. As a result, data must be both available and comparable across geographies which is not the case with the existing hedonic methods. However, the new autoregressive index, including only location information, can be easily applied across a diverse set of geographies. Furthermore, incorporating single sales enables the autoregressive method to improve the precision of both model parameters and price predictions.

3 Data

The available data contains sale prices for single family homes sold between July 1985 and September 2004 in twenty US metropolitan areas. 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. All sales in our data qualified for a conventional mortgage. Consequently, the data does not include very expensive homes, homes bought at subprime rates, or those bought solely with cash. While the data comprise a well-defined subset (and majority) of all home sales, the omissions can introduce some sample selection bias into all of the methods discussed in this paper.

Note that none of our analyses involve deleting repeat sales of significantly renovated homes as recommended by Case and Shiller as 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 type indices discussed in this paper. Summary counts of the number of sales and unique houses in the data are provided in Table 1.

For the analyses in Sections 5 and 6, we divide the data described above into training and test sets. Each model is applied to the training data; the test data are used for prediction (see Section 6). The test set contains the final sale of homes which sell three or more times in the sample period. In addition, for homes which 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 1 provide the size of the training and test sets. Given the size of the data set, it is impractical to provide all results in the current paper. Therefore, we only show results for a few cities which represent typical findings.

We include single sales homes in the training data since they are used in estimating the autoregressive model parameters. We do not include single sale houses in the test set since

Table 1: House and sale counts

City	Sales	Houses	Single Sales	Training Pairs	Test Set
Ann Arbor, MI	68,684	48,522	32,458	10,431	9,731
Atlanta, GA	376,082	260,703	166,646	59,222	56,127
Chicago, IL	688,468	483,581	319,340	105,708	99,179
Columbia, SC	7,034	4,321	2,303	1,426	1,287
Columbus, OH	162,716	109,388	67,926	27,601	25,727
Kansas City, MO	123,441	90,504	62,489	16,705	16,232
Lexington, KY	38,534	26,630	16,891	6,075	5,829
Los Angeles, CA	543,071	395,061	272,258	75,660	72,350
Madison, WI	50,589	35,635	23,685	7,714	7,240
Memphis, TN	55,370	37,352	23,033	9,372	8,646
Minneapolis, MN	330,162	240,270	166,811	46,206	43,686
Orlando, FL	104,853	72,976	45,966	16,147	15,730
Philadelphia, PA	402,935	280,272	179,107	63,082	59,581
Phoenix, AZ	180,745	129,993	87,249	25,830	24,922
Pittsburgh, PA	104,544	73,871	48,618	15,891	14,782
Raleigh, NC	100,180	68,306	42,545	16,372	15,502
San Francisco, CA	73,598	59,416	46,959	7,111	7,071
Seattle, WA	253,227	182,770	124,672	35,971	34,486
Sioux Falls, SD	12,439	8,974	6,117	1,781	1,684
Stamford, CT	14,602	11,128	8,200	1,774	1,700

none of the traditional repeat sales methods provides predictions for the price of such homes. (The autoregressive method can predict single sales homes, albeit not very accurately since the prediction would be based only on the weakly informative geographic indicator.)

4 Model Descriptions

In this section, we outline the original repeat sales index proposed by Bailey, Muth, and Nourse in 1963 (BMN) and two indices which are based on it: the original Case and Shiller index (C-S) and the FHFA HPI index (FHFA). These three indices are similar in and are all fit on the log price scale. Next, we describe the Standard and Poor's Case-Shiller based index (S&P/C-S) index which differs in that it is fit on the price scale and is based on Shiller

(1991). Finally, we end with a description of the autoregressive (AR) index which, while it makes use of the repeat sales concept, approaches index construction from a different perspective.

4.1 Bailey, Muth, and Nourse and two related indices

All three methods discussed next, BMN, C-S, and FHFA, are built on a model where the expected difference in log prices for two sales of a house is equal to the difference in the corresponding log indices along with a random error term. What differs is the error structure.

In particular, let there be $T+1$ time periods where sales can occur from 0, 1 $,\ldots,T$ and t be the subscript for time period. Using the BMN (1963) notation, for a pair of sales of a given house i, prices and indices are related by the following expression:

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

where P_{it} is the sale price of the ith house at the tth time period. For a pair of sales, t is the time at the first sale and t' the time at the second ($t' > t$). Finally, let B_t denote the general house price index at time t and $U_{itt'}$ the multiplicative error term for the sale pair. The model is fit on the logarithmic scale:

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

where p, b, and u are simply the logarithmic versions of the terms in (1). The model in (2) is fit using linear regression and the estimated log indices are converted into price indices using the exponential function. Note that only houses which have been sold twice are used to calculate the index–the remaining observations are omitted.

In the BMN model, the error term $U_{itt'}$ is assumed to have a log-normal distribution:

$\log\, U_{itt'} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_u^2\right)$ where *iid* denotes independent and identically distributed [1, p. 934]. Therefore the error variance is constant in the BMN method.

The C-S and FHFA methods both assume that the error term is heteroscedastic arguing that the length of time between sales should increase the variance of the log price differences. The C-S (1987, 1989) method includes three components to the log error term: individual contributions from each sale $(\sigma_u^2 + \sigma_u^2)$ and a random walk representing the time periods between sales $((t - t')\sigma_v^2)$. The resulting log error term has variance $2\sigma_u^2 + (t' - t)\sigma_v^2$. On the other hand, the log error term in the FHFA method includes only a random walk to reduce the chance of estimating negative weights during the fitting process which is possible in the C-S method. (This occurred in our analysis in Section 6.1 for the S&P/C-S method.) The log error variance as described in Calhoun (1996) is $(t' - t)\left(E\left[v^2\right] - E\left[vv'\right]\right) + (t' - t)^2 E\left[vv'\right]$ where v and v' are arbitrary steps in a random walk and $E[\cdot]$ is the expectation function.

While the error structures differ, both follow the same fitting procedure: (a) estimate the log index using regression (stopping here results in the BMN index), (b) estimate the variances of the log error terms by using the residuals from the first step, and (c) estimate the log index again using the weights computed by taking the reciprocal of the estimated standard deviations of the log error terms from step (b). Accordingly, the larger the gap time between sales, the lower the weights of the sale pairs in the final regression.

Note here that the reciprocal of the standard deviation estimates are used as weights; however the standard generalized least squares procedure uses the reciprocal of the estimated error *variances* not standard deviations as weights. Consequently, the C-S and FHFA indices are unbiased but do not have the lowest possible variance. This is undesirable especially if the regression estimates are to be used for prediction or for constructing confidence intervals.

4.2 S&P/Case-Shiller method

A variation of the 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 (2009). The S&P/C-S index is published for 20 Metropolitan Statistical Areas (MSA) and nationally. Unlike the indices described in Sec. 4.1, the S&P/C-S index is constructed from sale *prices* as opposed to log sale price differences. As a result it is considered an arithmetic index rather than a geometric index. However, the S&P/C-S index is fit following a similar three-step procedure as the C-S and FHFA indices but does incorporate instrumental variables.

As before, we have $T+1$ time periods from 0, 1, $\ldots, T$. For house i:

$$\begin{aligned} P_{i0} &= \beta_{t'} P_{it'} + U_{i0t'} && \text{first sale at time } 0, \\ 0 &= \beta_{t'} P_{it'} - \beta_t P_{it} + U_{itt'} && \text{first sale at time } t > 0 \end{aligned} \tag{3}$$

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_u^2 and σ_v^2 are the same variances from the original Case-Shiller index. The price index is $B_t = \frac{1}{\beta_t}$.

The response vector in (3) contains mostly zeros as the vast majority of sales do not occur in the base time period ($t = 0$). For those that do, note that the model is structured so that *future* sales are used to explain a *preceding* sale which is not intuitive.

The most important issue to note here is the error structure. The original C-S method specifies the error variance on the log scale to be $2\sigma_u^2 + (t'-t)\sigma_v^2$. The S&P/C-S index has the same model error variance despite the error being on the price scale and the index not being constructed from the differences of prices (or log prices). This error structure, as a result, does not follow directly from the BMN model setup and is imposed arbitrarily making interpretation difficult; Meissner and Satchell (2007) observe this inconsistency as well.

4.3 Autoregressive model

The AR method, proposed in Nagaraja, Brown, and Zhao (2011), is a variant of the hybrid index proposed by Case and Quigley (1991): it contains only ZIP code as a hedonic variable and incorporates single sales. However, there are some key differences. An alternate approach to house price modeling is to consider a sale not as one of an isolated pair but rather as one of a series of sales. The AR model considers all sales of the same house as components of one series. Therefore, the entire data is comprised of thousands of short series, one series for each house. For single sales, the series has a length of one.

In theory, a house has a price at each time period. However, the price is observed only when the house is sold. These are the prices in the data set. Consequently, we can think of each house price series as an autoregressive process which is observed when a sale occurs. Single and repeat sales are incorporated naturally under this setup; furthermore, gap times are simply the periods where house prices are not observed.

This log price model contains three components: (a) a log time effect, (b) weak hedonic information through the use of ZIP code as a proxy for location, modeled using a random effects term, and (c) an underlying stationary, autoregressive process which handles the serial nature of house sales. The log time effect, which is converted into an index by exponentiating and dividing by the base period value, is comprised of information from both single and repeat sales homes, the latter receiving a much higher weight because more information is known about repeat sales houses through having multiple, observed prices.

In particular, let $p_{i,j,k}$ be the log price of the jth sale of the ith house in ZIP code k. We define the notation $t(i,j,k)$ to indicate the time period t where the jth sale of the ith house in ZIP code k is sold. Let $\mu + \beta_{t(i,j,k)}$ be the log time effect for time period, $t(i,j,k)$, and let $\gamma(i,j,k)$ be the gap time, or $t(i,j,k) - t(i,j-1,k)$, if it is the second or higher sale. Finally, ZIP code is modeled as a random effect: $\tau_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_\tau^2)$. Then,

$$\begin{aligned} p_{i,1,k} &= \mu + \beta_{t(i,1,k)} + \tau_k + \varepsilon_{i,1,k} & j = 1 \\ p_{i,j,k} &= \mu + \beta_{t(i,j,k)} + \tau_k + \phi^{\gamma(i,j,k)} \left(p_{i,j-1,k} - \mu - \beta_{t(i,j-1,k)} - \tau_k \right) + \varepsilon_{i,j,k} & j > 1 \end{aligned} \tag{4}$$

The error distributions are as follows: $\varepsilon_{i,1,k} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_{i,j,k} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j,k)}\right)}{1-\phi^2}\right)$, and all $\varepsilon_{i,j,k}$ are assumed to be independent. The adjusted log price value, $p_{i,j,k} - \mu - \beta_{t(i,j,k)} - \tau_k$, is a stationary first order autoregressive time series (AR(1)) with autoregressive coefficient ϕ where $|\phi| < 1$. Maximum likelihood estimation is used to fit this model. Details can be found in Nagaraja, et al (2011).

Intuitively, we would expect that the previous sale price of a house is less valuable the larger the gap time. The AR model naturally produces this feature in two keys ways. First, as gap time increases, the correlation between the adjusted log prices between sale pairs (however you want to define them) decreases. Second, an indirect effect, as gap time increases, the error variance increases in the model (see the expression for $\varepsilon_{i,j,k}$).

Furthermore, the error variance is much larger for single sales ($\varepsilon_{i,1,k}$) as opposed to repeat sales ($\varepsilon_{i,j,k}$) as less information is known about the former. As a result, single sales are less influential in the model estimation than repeat sales as influence is inversely proportional to the variance.

5 Qualitative Comparisons

There are five main conceptual differences among the indices outlined in the previous section: (a) single and repeat sales, (b) effect of gap time on sale prices, (c) use of hedonic information, (d) treatment of more than two sales, and (e) geometric versus arithmetic index construction. We disucss each of these differences in the following section.

5.1 Single and repeat sales

There are two consequences of excluding single sales. First, to have sufficient data to construct a stable index, larger geographical areas may be required. While data spanning a longer period will result in a higher number of repeat sales, the number of newly built houses also increases. Unless the geographical area in question is extremely stable such as the Herengracht canal area in Amsterdam studied by Eichholtz (1997), the proportion of repeat sales among all house sales does not increase as fast as one might expect.

Second, repeat sales homes may be fundamentally different from single sales homes, resulting in biased indices as shown in Case, Pollakowski, and Wachter (1991). If true, then indices derived only from repeat sales homes are indicative of changes in such homes at best, not the entire housing market. It has been hypothesized that a higher proportion of repeat sales are "starter homes" where young families live which are soon traded for larger, costlier, and nicer homes after only a few years (Clapp, et al 1991, p. 271). Meese and Wallace (1997, p. 55) do find significant differences while Clapp, et al (1991) have inconclusive results. The autoregressive method attempts to take a middle ground on this issue by including both single and repeat sales, placing more emphasis on the repeat sales information.

5.2 Effect of gap time on sale prices

The role of gap time is an important issue for each of the five indices. Intuitively, one would expect the time between sales to affect the usefulness of the previous sale price. The BMN method is the only technique which does not include this feature. The C-S, FHFA, and S&P/C-S based methods incorporate gap time into the error structure, but in very different ways. Finally, the AR method includes gap time in the error structure as an inherent statistical consequence of the underlying autoregressive component in the model. Table 2 lists the theoretical error variance for each method where γ is the gap time between sales.

In order to see whether each method captures the respective error variance, we compare the estimated variance and gap time relationship with the empirical data. We show results for two typical cities: Orlando, FL and Philadelphia, PA. In each plot in Figure 1 and Figure 2, the expected variance given a gap time is plotted (line) and the variance of the training set residuals for each gap time are plotted (points). These residuals (residual=observed price − predicted price) are computed from predicting the second sale of sale pairs in the training set from each model. The former represents the theoretical variance (from Table 2) and the latter the empirical variance. Some gap times contain very few sale pairs; those with 15 or fewer observations are indicated with an "x". Note that only the S&P/C-S results are on the price scale as it is the only method fit on the price scale.

From these plots we can determine whether the theoretical error variance, as specified by each model, matches the empirical results for each city. It is clear that the error variance is indeed dependent on gap time and the BMN method performs worst in this regard. The FHFA supposes a parabolic relationship which is not entirely supported by the empirical results. The C-S and S&P/C-S methods seem to be the best at capturing the error variance. The AR method theoretical error variance seems to follow the empirical error variance better for Orlando, FL than for Philadelphia, PA; the remaining cities show varied results as well.

For the AR method only, the gap time has a second role: the ϕ^γ term in (4). As the gap time increases, the correlation between adjusted log sale prices decreases following ϕ^γ. This feature has a larger effect, and therefore, more important function, than the gap time and error variance relationship. For more details, refer to Nagaraja, et al (2011).

5.3 Hedonic information

In practice, repeat sales methods are popular in part because they require little information about each sale: time of sale, price, and a unique house identifier. However, for these data

Figure 1: Variance by gap time for Orlando, FL

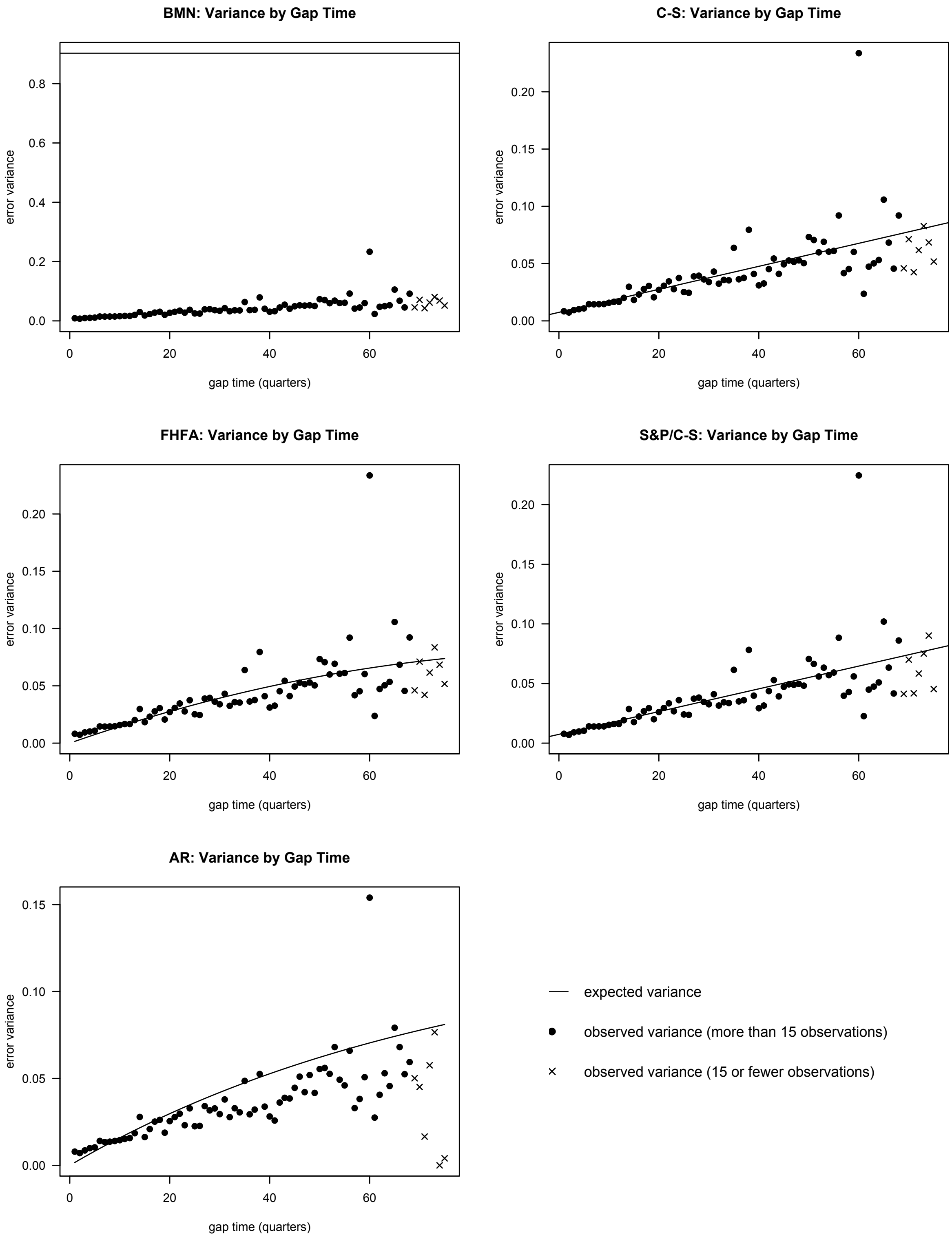


Figure 2: Variance by gap time for Philadelphia, PA

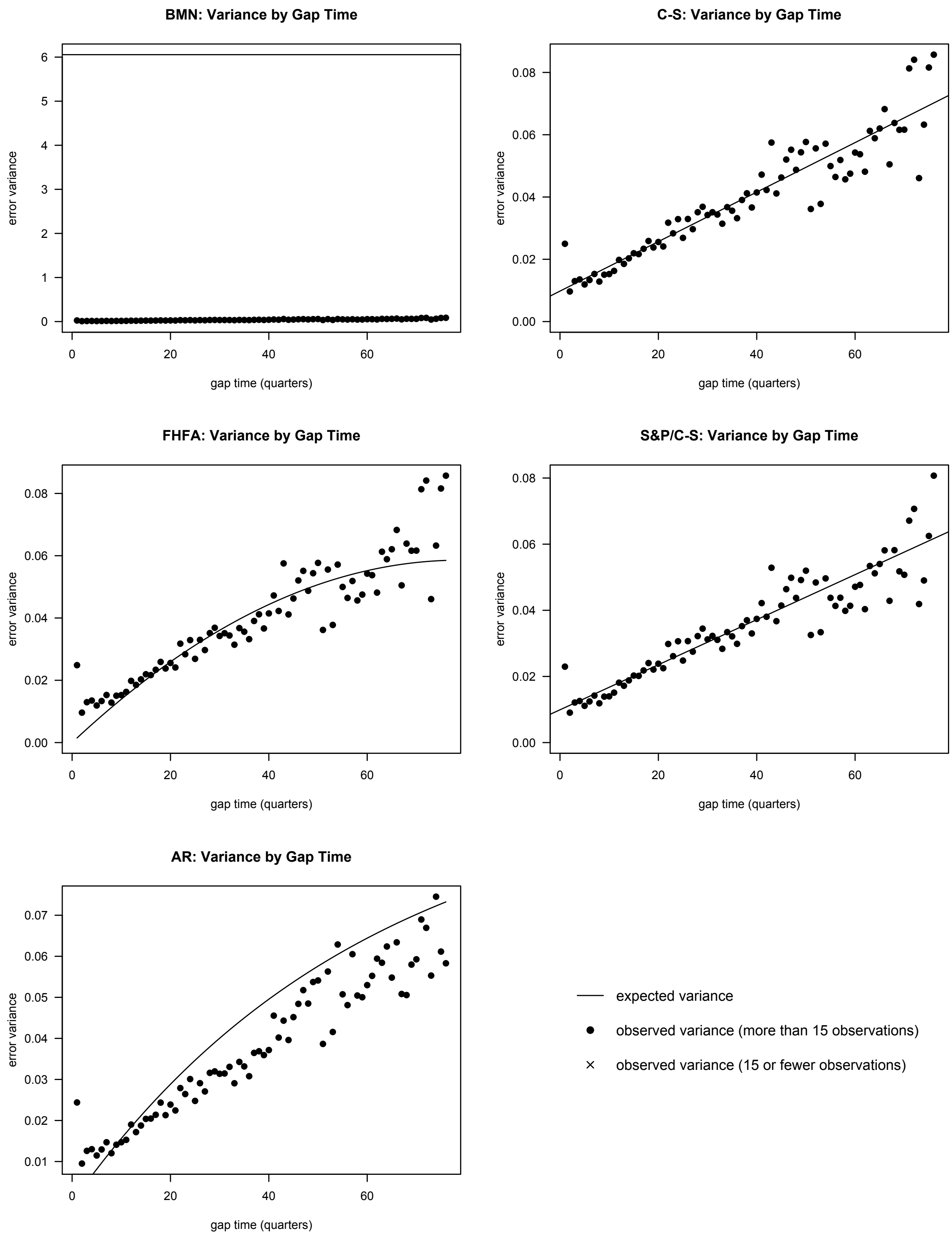


Table 2: Error variance in the fitted model (γ is gap time)

Index	Error Variance
BMN	σ^2
C-S	$2\sigma_u^2 + \gamma\sigma_v^2$
FHFA	$\gamma\left(E[v^2] - E[vv']\right) + \gamma^2 E[vv']$
S&P/C-S	$2\sigma_u^2 + \gamma\sigma_v^2$
AR	$\frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma}\right)}{1-\phi^2}$

to be sufficient, we must make an assumption: only sales of an *identical* house should be compared. There are a few ramifications of this strong assumption. As Englund, et al (1999) discuss, houses which have significantly improved or deteriorated between sales should be removed as they violate the equivalence property. Obtaining quality data on these changes can be difficult. Palmquist (1982) proposes a depreciation factor to account for age. This factor could ostensibly be applied to any of the five methods discussed in this paper. However, any depreciation method used requires additional information about the property, and this is often unavailable.

Finally, it is implicitly assumed within the equivalence property that the previous sale price is an adequate proxy for hedonic characteristics. Therefore, no additional information about the house is required and the influence of the hedonic characteristics must not change over time. There is some evidence that this may not be the case. Case and Quigley (1991) propose a hybrid model which shows the utility of incorporating hedonic variables in repeat sales models (in addition to including single sales and homes which have changed significantly). Gillen, Thibodeau, and Wachter (2001) show that single-family house prices are spatially autocorrelated in Philadelphia, PA. The AR model adds a location effect through ZIP code and it acts as a useful hedonic variable (see the end of Section 6.1 for evidence).

5.4 Treatment of more than two sales

The BMN, C-S, FHFA, and S&P/C-S methods are designed for homes that sell only twice. Little consideration is given for homes which sell more than twice and how to construct appropriate sale pairs. For example, if there are three sales, pairs can be constructed as (first, second) and (second, third). Alternatively, we could construct (first, second), (first, third), and (second, third). In this paper, we consider only the first pair construction method.

Under those conditions, Bailey, et al (1963) suggest adding a fixed effect for property or computing a weighted regression where the impact of each sale pair with a sale which appears in another sale pair is split across sale pairs. In the C-S method, the covariance across sale pairs with common sales is $Cov(p_{it'} - p_{it}, p_{it''} - p_{it'}) = -\sigma_u^2$ which can be incorporated into the weight matrix; the S&P/C-S method, with the identical error structure, would have a similar setup. In the FHFA method, the correlated random walk steps create a very complex relationship across sales. In particular, $Cov(p_{it'} - p_{it}, p_{it''-p_{it'}}) = (t' - t)(t'' - t')\zeta$ where $\zeta = E[vv']$. Technically, these correlation structures can be incorporated into the respective models even though none of the methodologies do so. As three or more sales of a single house is relatively uncommon, in data sets like ours, this additional step can be ignored with minimal loss. On the other hand, the AR method has a coherent treatment of multiple sales. This is because the model handles homes not as components of sale pairs but rather as a single series of sales. As a result, there is no need for any adjustments.

5.5 Geometric versus arithmetic indices

The S&P/C-S index is the only method which is an arithmetic index. It is fit on the price scale whereas the BMN, C-S, FHFA, and AR indices are all fit on the log price scale and are therefore geometric indices. That is, when the models are transformed from the log price scale to the price scale, the additive model (on the logarithmic scale), becomes a multiplicative model (on the price scale). Shiller (1991), Goetzmann (1992), and Goetzmann and Peng (2002) all argue that arithmetic house price indices are more appropriate, and perhaps less biased, than geometric indices. They have two main reasons for this claim.

First, they suggest that changes in prices are more easily interpreted as changes in dollar amounts as opposed to percentage changes (logarithmic). Second, by virtue of being fit on the price scale, arithmetic indices are what Shiller (1991) calls "value-weighted." This concept could be relevant if more expensive homes behaved differently than less expensive ones (p. 110). Geometric indices treat all houses roughly the same in terms of log price (but not necessarily by gap time). If different house price levels should be treated differently, then it may be better to construct separate indices.

6 Quantitative and Index Comparisons

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

6.1 Prediction results

All five indices are constructed from base models for house price; therefore, prediction of individual house prices follows quite naturally as an evaluation technique. Note that 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 a subsample of the data not used to fit the model. In essence we are predicting prices within the scope of the data set, rather than forecasting prices at times outside this time span.

Each method is fit using the training data set. The sale price for each house in the test set is then predicted. For the methods fit on the log price scale (BMN, C-S, FHFA, AR), the log price prediction is converted to the price scale as explained in Nagaraja, et al (2011). 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 in the test set, $\hat{P}_i$ denotes its predicted price, and n is the number of observations in the test set. The results are calculated and reported separately for each of the 20 metropolitan areas in Table 3.

Note, 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 calculated. Recall, that this is the type of problem that the FHFA method tries to avoid.

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? The answer to this question may be of interest even when an index is the overall end goal. Predictions, however, also provide a better quantitative measure of the effectiveness of the index in describing market trends. Using this metric, we find 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 result in only minor changes to the RMSEs. However, the AR index performs best.

Components of the AR model

The AR method has three components: (a) the autoregressive process, (b) inclusion of ZIP code, and (c) use of single sales. Feature (a) combined with the log time effect form the basis of the model. On the other hand, components (b) and (c) have been added to the fundamental model. We fit three additional versions of the AR model to explore these components: AR model without ZIP code (column 7 in Table 3), AR model without ZIP code fit only on repeat sales data (column 8 in Table 3), and the AR model with ZIP code fit only on repeat sales data (column 9 in Table 3). In the latter two versions, the single sales are removed from the training data set and so are not used in the model fitting process. (Note that in the final version, a negligible number of observations in the test set have to be dropped because they come from ZIP codes not represented in the training set.)

Traditional repeat sales methods assume that no hedonic information is needed; if true, ZIP code should be superfluous to the model. In this regard, we should compare the results from column 6 and column 7: the original AR model and the AR model without the ZIP code effect component. While the basic model still has better predictions than BMN, C-S, S&P/C-S and FHFA methods, the improvement is much less dramatic than in the AR model including ZIP code. We conclude that the location effect is an integral part of the model.

Finally, if repeat sales homes are generally similar to single sales homes, then including them should improve the prediction of repeat sales home prices in the test set. However, we find the opposite to be true. In nearly all cases, excluding single sales from the training set improved the RMSE value. We can see this in Table 3 when comparing column 6 with 9 (AR including ZIP code) or if we compare columns 7 and 8 (AR excluding ZIP code). This is confirmation of a difference between single and repeats sales homes. While, one may argue that the AR model should be fit excluding the single sales because of improved prediction, we feel it is more important to model the visible housing market. Furthermore, the inclusion of single sales, like in the Case and Quigley (1991) hybrid model, increase the apparent precision of both the index estimates and the predicted prices. This is because, by including single sales, the sample size used to fit the model increases considerably and thus decreases the standard error of estimates.

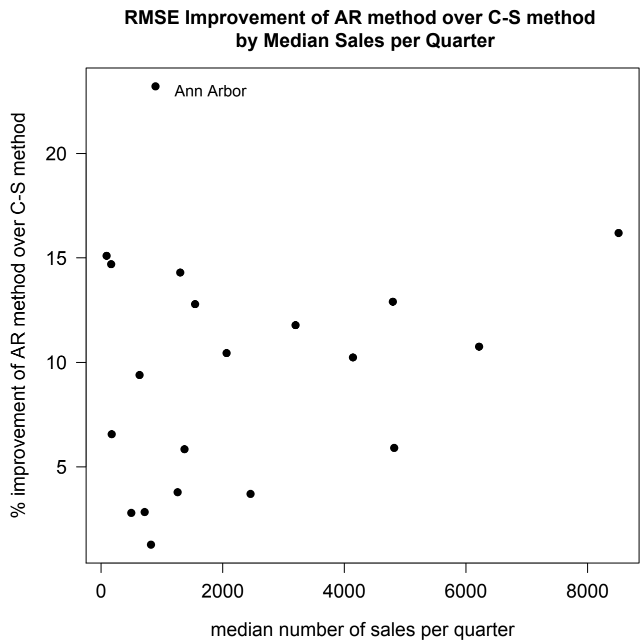
RMSE versus sample size

One might expect that the improvements in RMSE for the AR method over the traditional repeat sales methods would be higher for smaller cities which have fewer total sales. In Figure 3, we plot the percent RMSE improvement of the AR method over the C-S method for each city against the median number of sales per quarter for the corresponding city (as a proxy for city size). We find, however, for the AR method there is no discernible pattern in the RMSE improvement according to city size other than that results are more variable for smaller cities which is expected. The one outlier is Ann Arbor, MI (labeled on the graph) which is one of the smallest cities in the data but has the largest RMSE improvement. Qualitatively similar results are obtained if the C-S index is replaced with the BMN, S&P/C-S, or FHFA index.

Table 3: Test set RMSE (in dollars)

Metropolitan Area	BMN	C-S	S&P/C-S	FHFA	AR	AR (no ZIP)	AR (no ZIP, no single sales)	AR (ZIP, no single sales)
Ann Arbor, MI	53,709	53,914	52,718	54,024	**41,401**	44,362	42,639	39,825
Atlanta, GA	35,456	35,494	35,482	35,503	**30,914**	33,977	33,595	29,002
Chicago, IL	42,923	42,960	42,865	42,976	**36,004**	39,202	38,910	34,623
Columbia, SC	42,207	42,263	42,301	42,290	**35,881**	36,377	35,136	34,789
Columbus, OH	30,550	30,543	30,208	30,545	**27,353**	28,126	27,656	26,196
Kansas City, MO	27,682	27,724	–	27,730	**24,179**	24,964	24,819	23,164
Lexington, KY	21,748	21,740	21,731	21,741	**21,132**	21,501	21,602	20,832
Los Angeles, CA	41,918	41,949	41,951	41,959	**37,438**	41,006	40,856	37,567
Madison, WI	30,979	30,942	30,640	30,950	**28,035**	28,687	28,464	27,317
Memphis, TN	25,311	25,306	25,267	25,311	**24,588**	25,069	24,782	23,613
Minneapolis, MN	35,402	35,538	34,787	35,565	**31,900**	33,233	31,904	30,157
Orlando, FL	30,187	30,215	30,158	30,228	**28,449**	29,317	29,256	27,287
Philadelphia, PA	35,308	35,333	35,350	35,338	**33,246**	34,736	34,502	32,878
Phoenix, AZ	29,295	29,334	29,350	293,56	**28,247**	30,232	28,832	26,687
Pittsburgh, PA	30,732	30,812	30,135	30,858	**26,406**	26,508	26,483	26,058
Raleigh, NC	26,873	26,856	26,775	26,855	**25,839**	26,564	26,864	25,157
San Francisco, CA	50,513	50,573	50,249	50,499	**49,927**	50,778	51,347	50,488
Seattle, WA	43,533	43,606	43,486	43,631	**38,469**	42,330	41,338	36,290
Sioux Falls, SD	21,527	21,576	21,577	21,525	**20,160**	20,190	20,398	20,334
Stamford, CT	67,661	67,668	68,132	67,579	**57,722**	61,805	62,688	59,027

Figure 3: RMSE improvement against sample size



6.2 Index comparison

Despite the differences in methodology, the traditional repeat sales and the autoregressive index track each other exceptionally well at the macro level (the indices were computed using the training set data). For comparison purposes, we include the median price index. The correlation between each pair of log index return series is 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 (except for the median index); however, if we plot the indices for three metropolitan areas Atlanta, GA, Minneapolis, MN, and Pittsburgh, PA as in Figures 6-8, we can see that the actual values of each index differ. These cities were chosen to represent the range of results. The plot on the left shows the index produced from each method; on the right is the index at time t subtracted from the average index level at time t (average of the 6 indices). From this plot, differences among the indices can be more

Table 4: Correlation Among Log Index Returns for Minneapolis, MN

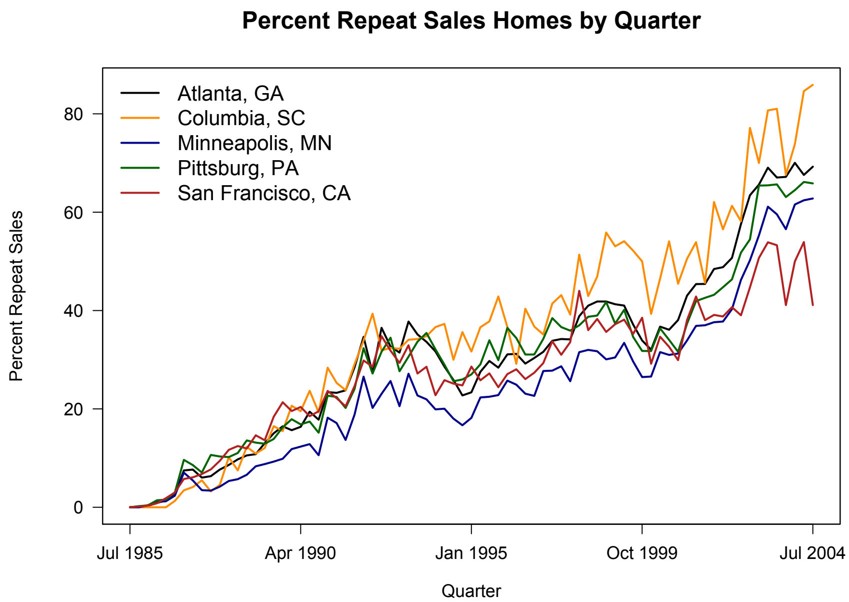
	BMN	C-S	S&P/C-S	FHFA	AR	Median
BMN	1	0.9778572	0.9558543	0.9654379	0.8805295	0.4153559
C-S	–	1	0.9778387	0.9984017	0.9092791	0.3717965
S&P/C-S	–	–	1	0.9753913	0.8781704	0.3894031
FHFA	–	–	–	1	0.9130247	0.3578883
AR	–	–	–	–	1	0.6023298
Median	–	–	–	–	–	1

easily detected. While we only chose these three cities to show, we did obtain similar results for the remaining cities.

The BMN, C-S, and FHFA indices are nearly all the same; this is not surprising as the methods applied are similar. The median, S&P/C-S, and AR, indices tend to differ from the others, but not in any systematic manner. 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. In contrast, the repeat sales indices include no single sales so only repeat sales information is used. 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.

The fact that none of the indices is consistently higher or lower than the others could possibly reflect varying growth rates across the cities. In Figure 4 the percentage of repeat sales homes in each quarter is plotted for a selection of cities. Note that for this plot only, in any given quarter a house is considered a repeat sale only if it was sold at least once *before*. 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 rather than a more even mix of new homes and homes which have sold only once 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

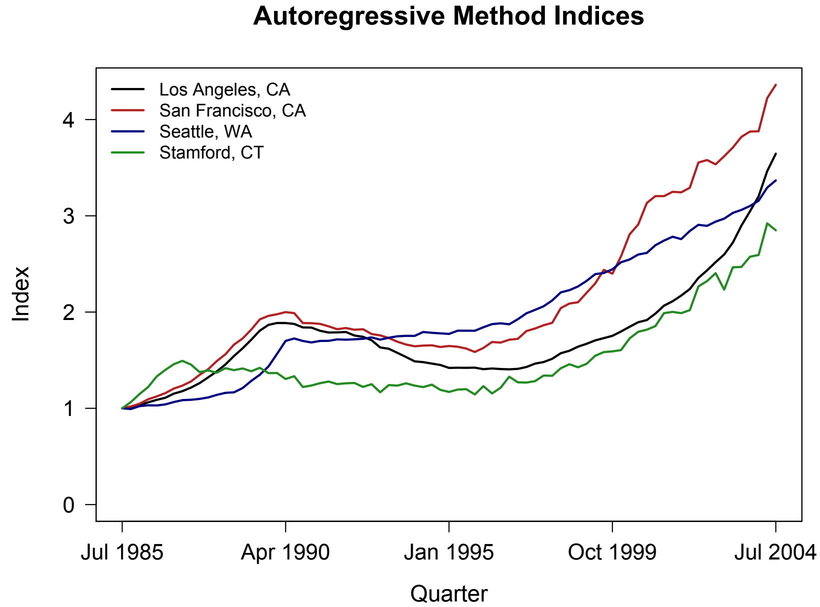
Figure 4: Percentage of repeat sales by quarter for a selection of cities



for Columbia, SC, at nearly 86%. However, in our data, we cannot distinguish between new homes and singles sales of old homes. Therefore, we cannot determine how differences in growth rates across cities affect the indices.

The sample period ends in September 2004 meaning that the recent housing crisis is not included in this analysis. However, there are four areas where the AR housing index does decrease significantly during the sample period: Los Angeles, CA, San Francisco, CA, Seattle, WA, and Stamford, CT. These indices are plotted together in Figure 5. Apart from Los Angeles, CA, which is described in Nagaraja, et al (2011) in more detail, the AR model is a good fit for the data for the remaining three cities. Therefore, it is possible that the AR model can be appropriate when housing markets decline, assuming the model is a good fit generally.

Figure 5: Metropolitan areas where the AR index decreases in sample period



7 Summary

The five indices, BMN, Case-Shiller, FHFA, S&P/Case-Shiller, and the autoregressive index are all based upon the repeat sales idea. We have also proposed the use of individual price prediction as a useful metric to compare indices. Indices are used for local applications, so price prediction can be useful. Furthermore, all other existing metrics are qualitative and often very difficult or nearly impossible to test on available data. Prediction, however, is straightforward and objective and therefore, we feel, an important part of the index assessment process. All of the methods, other than the BMN index, incorporate adjustments for gap time; however, only the autoregressive method also includes single sales and hedonic information in the form of a ZIP code effect. The latter feature has been shown to be very important for predictive power.

The question we are trying to explore in this paper is how to tell if a house price index is informative. If usability is key, all of these indices are adequate–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 traditional repeat sales indices can truly represent the housing market. While all of the indices (including the median index) exclude houses that do not sell, the median and autoregressive index 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 and is a statistical model. However, none of these standards indicates whether an index is truly measuring what it is supposed to. We feel the best yardstick in this regard is predictive ability. In this case, the autoregressive index is the clear winner since for *all* twenty cities, the RMSE values were the lowest among all of the indices. In fact, the autoregressive model seems to best embody what an index should represent.

Figure 6: Indices for Atlanta, GA

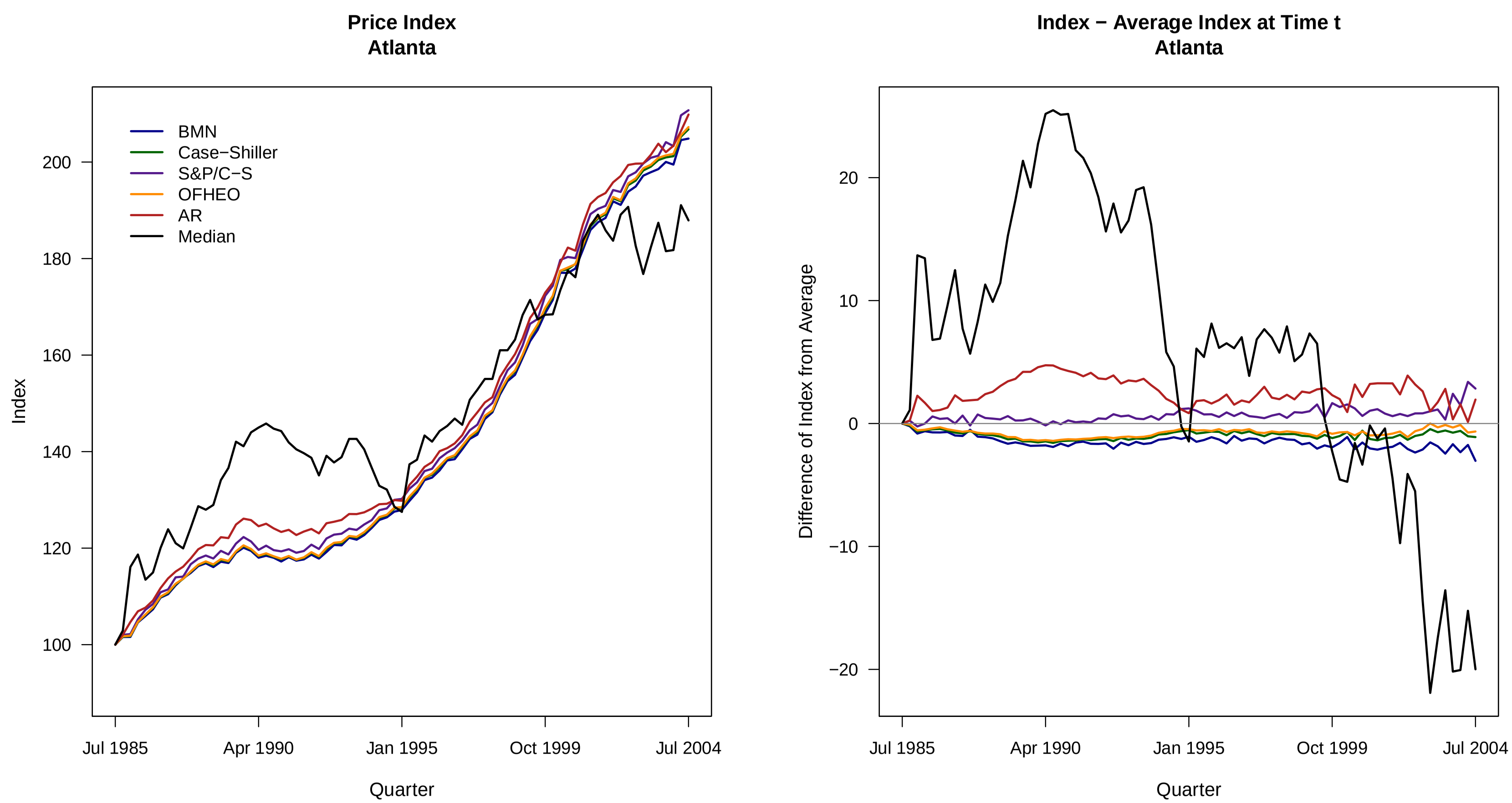
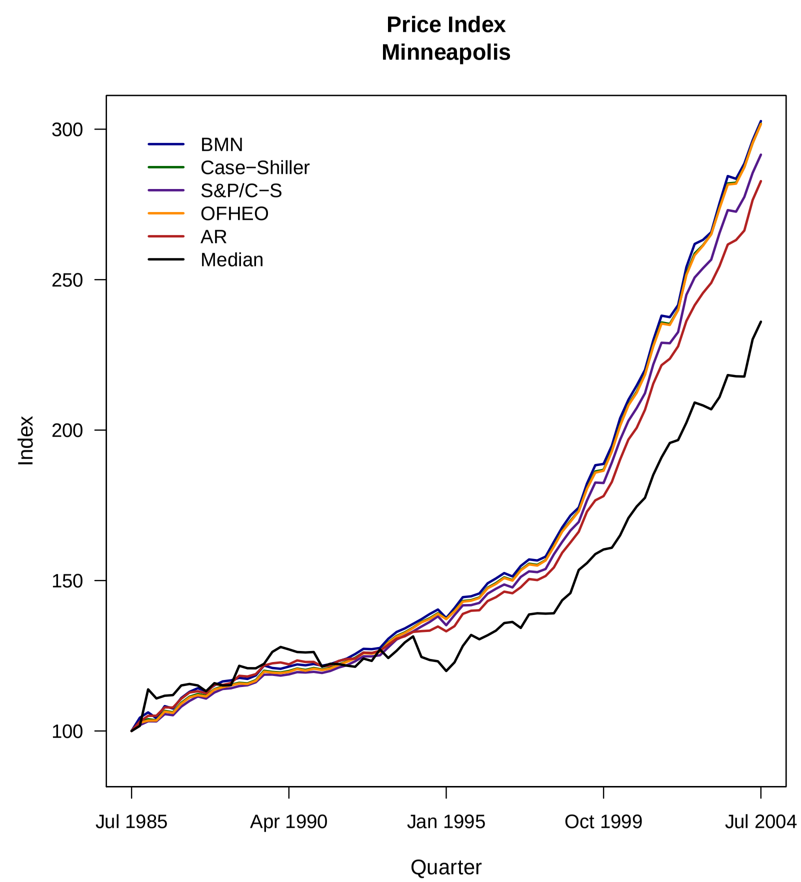


Figure 7: Indices for Minneapolis, MN



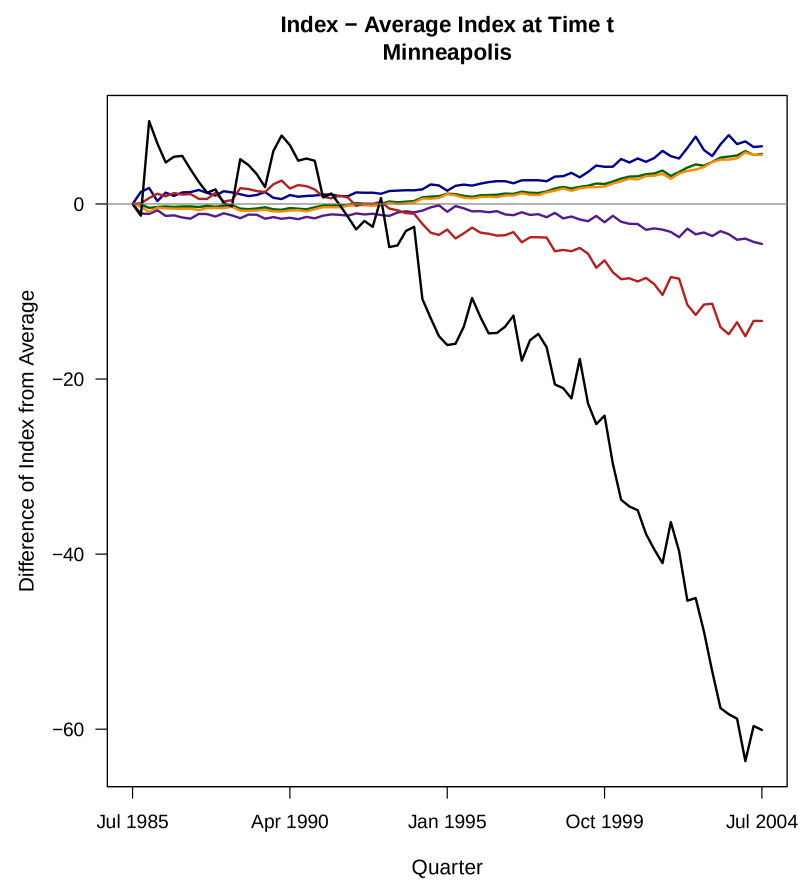
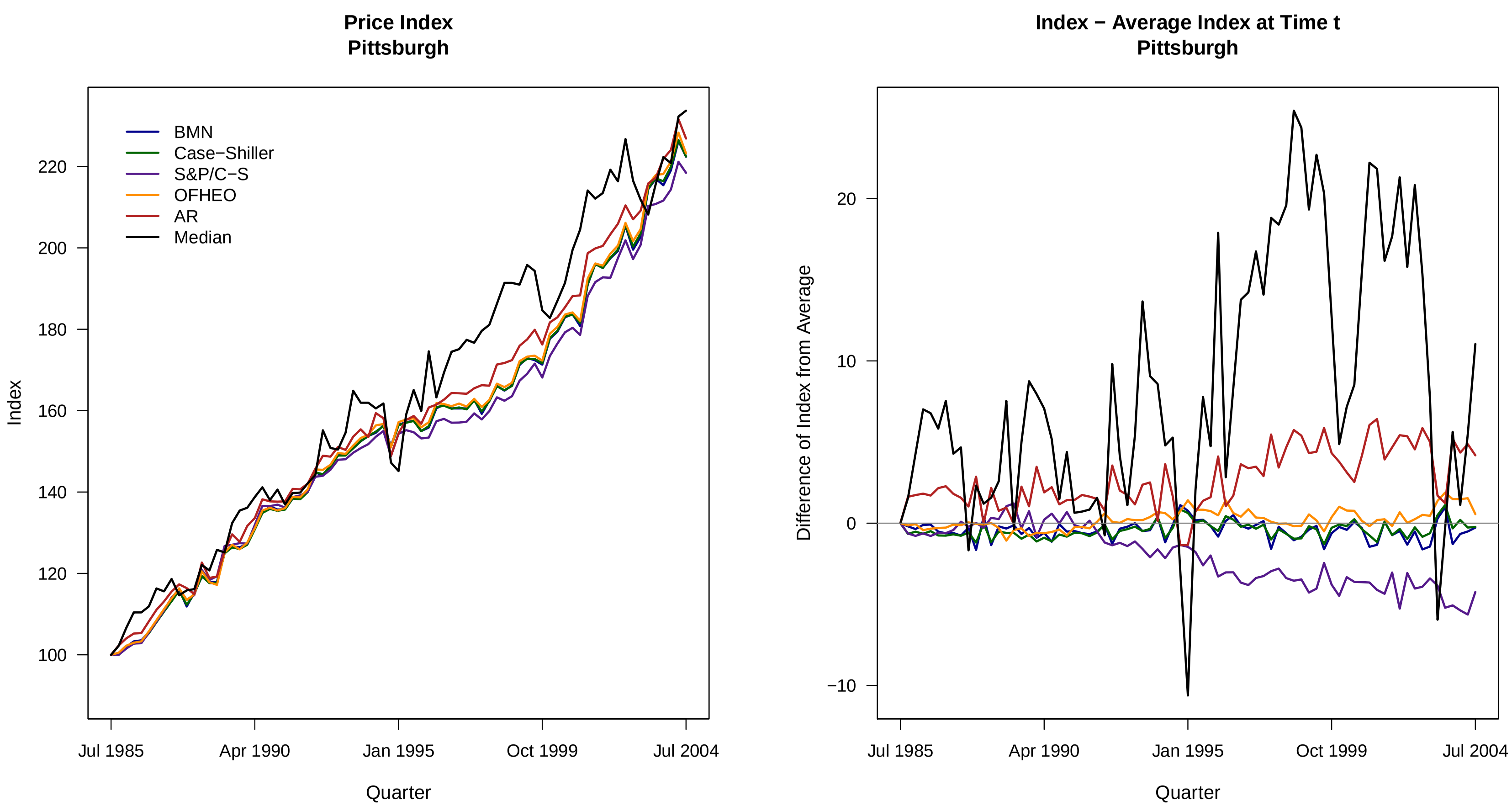


Figure 8: Indices for Pittsburgh, PA



Acknowledgements

We would like to thank Linda H. Zhao from The Wharton School, University of Pennsylvania for providing the data for our analysis and to the anonymous referees who provided helpful comments. Susan Wachter thanks the Research Sponsor Program of the Zell/Lurie Real Estate Center at the Wharton School of the University of Pennsylvania for its support. This work was funded in part by by a Faculty Research Grant from Fordham University.

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 [online]. Availble at http://www.fhfa.gov

[3] Cannaday, R.E., Munneke, H.J., Yang, T.T. (2005). A multivariate repeat-sales model for estimating house price indices. *Journal of Urban Economics.* **57** 320-342.

[4] 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.

[5] Case, B., Quigley, J.M. (1991). The dynamics of real estate prices. *The Review of Economics and Statistics.* **73** 50-58.

[6] 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.

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

[8] Chau, K.W., Wong, S.K., Yiu, C.Y. (2005). Adjusting for non-linear age effects in the repeat sales index. *The Journal of Real Estate Finance and Economics.* **31** 137-153.

[9] Clapp, J.M., Giacotto, C. (1992). Estimating price indices for residential property: a comparison of repeat sales and assessed value methods. *Journal of the American Statistical Association.* **87** 300-306.

[10] Clapp, J.M., Giacotto, C. (1998a). Price indices based on the hedonic repeat-sales method: application to the housing market. *Journal of Real Estate Finance and Economics.* **16** 5-26.

[11] Clapp, J.M., Giacotto, C. (1998b). Residential hedonic models: a rational expectations approach to age effects. *Journal of Urban Economics.* **44** 415-437.

[12] 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.

[13] Eichholtz, P. M.A. (1997). A long run house price index: the *Herengracht* index, 1628-1973. *Real Estate Economics.* **25** 175-192.

[14] Englund, P., Quigley, J.M., Redfearn, C.L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *Journal of Real Estate Finance and Economics.* **19** 91-112.

[15] Gatzlaff, D.H., Haurin, D.R. (1997). Sample selection bias and repeat-sales index estimates. *Journal of Real Estate Finance and Economics.* **14** 33-50.

[16] Gatzlaff, D.G., Ling, D.C. (1994). Measuring changes in local housing prices: an empirical investigation of alternative methodologies. *Journal of Urban Economics.* **35** 221-244.

[17] Gillen, K., Thibodeau, T., Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics.* **23** 5-30.

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

[19] Goetzmann, W.N., Peng, L. (2002). The bias of the RSR estimator and the accuracy of some alternatives. *Real Estate Economics.* **30** 13-39.

[20] Goodman, A.C., Thibodeau, T.G. (1996). Dwelling age heteroskedasticity in repeat sales house price equations. *Real Estate Economics.* **26** 151-171.

[21] Knight, J.R., Dombrow, J., Sirmans, C.F. (1995). A varying parameters approach to constructing house price indexes. *Real Estate Economics.* **23** 187-205.

[22] 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.

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

[24] Nagaraja, C.H., Brown, L.B., Zhao, L. (2011). An autoregressive approach to house price modeling. *Annals of Applied Statistics.* **5** 124-149.

[25] Nagaraja, C.H., Brown, L.B., Zhao, L. (2010). Supplement to "An autoregressive approach to house price modeling." DOI: 10.1214/10-AOAS380SUPP.

[26] Noland, C.W. (1979). Assessing hedonic indexes for housing. *The Journal of Financial and Quantitative Analysis.* **14** 783-800. (Proceedings of 14th Annual Conference of the Western Finance Association, June 21-23, 1979).

[27] Palmquist, R.B. (1979). Hedonic price and depreciation indexes for residential housing: a comment. *Journal of Urban Economics.* **6** 267-271.

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

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

[30] Standard and Poor's (Nov. 2009). S&P/Case-Shiller® Home Price Indices [online]. Available at http://www.standardandpoors.com

[31] Yeats, M.H. (1965). Some factors affecting the spatial distribution of Chicago land values, 1910-1960. *Economic Geography.* **41** 57-70.