Estimating house price growth with repeat sales data: What's the aim of the game?

Ferdinand T. Wang and Peter M. Zorn*

December 13, 1999

Abstract

Since the seminal work of M. Bailey, R. Muth, and H. Nourse (1963, *J. Amer. Statist. Assoc.* **58,** 933–942), numerous articles have been written about repeat sales and other methods for constructing house price indices. Our justification for producing yet another paper on this subject is to reemphasize fundamentals. We focus on the basic building blocks—asking questions about what the underlying target is, how repeat sales goes about estimating the target, and when a particular index might be used in practice—rather than on more complex, higher level concerns such as statistical or modeling accuracy. We find that much of the debate over index methodology can be distilled to implicit and largely unrecognized disagreement over the desired target or the intended application. Consequently, we contend that paying greater heed to fundamental questions offers significant rewards to both researchers and practitioners.

* Ferdinand T. Wang (ferdie_wang@freddiemac.com) is Sr. Economist and Peter M. Zorn (peter_zorn@freddiemac.com) is Director, Freddie Mac, Housing Economics and Financial Research, Mailstop #484, 8200 Jones Branch Drive, McLean, VA 22102-3110.

1 Introduction

Over the past decade researchers have modified and extended the basic method of Bailey et al. (1963) for using repeat sales data to compute house price indices and growth rates. Proponents of these methods have typically tried to motivate their competing strategies through either modeling or statistical arguments, with particular attention paid to differences in the conditioning variables or the error structure. In this article we take a more basic approach. We concentrate instead on considerations of the appropriate "target" and where procedures are "aiming." While not proposing the abandonment of model comparisons, we believe that this lower level focus offers significant benefits. Primary among these is the realization that much of the ongoing debate in the research community over house price index methodology is rooted in simple and easily understood issues.

We define the "target" to be the population measure that users of an index would like to hit regardless of the method. Focusing on the target explicitly introduces the practical question of what an index is supposed to do. In contrast to issues regarding the statistical properties of some estimator (such as bias and efficiency), this simple query is often quite easy to answer. We define the "aim" of an estimation procedure to be its output when applied to idealized population data. Note that while the target is entirely independent of the method, the aim is completely defined by the method. With this framework, we contend that it is crucial for house price index practitioners to *choose a method that aims at the desired target.* Achieving such an alignment is unlikely without a clear and explicit focus.

The use of idealized population data allows us to abstract away from real data problems. It highlights what any given method is "trying" to do, and alternative methods are trying to do different things. Recognition of this fact resolves much of the current brouhaha over competing house price indices. An additional benefit of using idealized data is the appreciation that many of the difficulties attributed to various methods (repeat sales as well as others) are actually due to characteristics of the sample. Explicitly identifying the aim of a procedure provides practitioners with a reliable metric to assess how well methods perform with real world data.

We have several secondary goals for this article that account for much of its length. We hope to simplify comparison across alternative repeat sales methods by providing a unified presentation and notation for measuring house price appreciation. We also try to give the intuition behind the key aspects of alternative models. Finally, we detail a number of potential problems with the repeat sales method that have been confused, mentioned only briefly, or overlooked entirely.

The organization of the article is as follows. We begin in Section 1 by describing the idealized population data and one very natural candidate for a target. In Section 2 we show what the repeat sales estimators are aiming for and what they actually yield under real world data. In Sections 3 and 4 we address problems that have been raised in the literature using our taxonomy of target and aiming difficulties. In Section 3 we show that many differences in proposed techniques reduce to distinctions in their implicit targets, and in Section 4 we concentrate on statistical issues (i.e. how well repeat sales estimators hit what they are aiming for). In Section 5 we briefly discuss alternative methods for measuring house price appreciation, and we conclude in Section 6.

1.1 Idealized population data and a natural target

Idealized population data are useful for determining both target and aim. By explicitly displaying every value in the population, they make identification of intuitively reasonable target parameters easy. This facilitates and promotes closer consideration of exactly what practitioners wish to capture quantitatively with index values or growth rates. Idealized data also abstract away real-world data problems, so that when comparing different methods attention can first be focused on the outputs of those methods rather than on the technical details of the procedures themselves. Idealized populations demonstrate what a method is trying to do with the data by defining the outcome of that method under the best possible circumstances. In providing a clear metric to assess statistical measures of accuracy such as bias and consistency, this should aid researchers in deciding if a method is appropriate and if it does a good job in practice.

With "perfect information" we would know every house's price at every time period. Specifying the price of house j at time t as $\pi_{j,t}$, and denoting the population size and the total number of time periods under consideration by the symbols N and T respectively, we may write out the population of prices in a "big matrix" as follows:

					time					
0	1	2	3	4	5	6	7	8	...	T
$\pi_{1,0}$	$\pi_{1,1}$	$\pi_{1,2}$	$\pi_{1,3}$	$\pi_{1,4}$	$\pi_{1,5}$	$\pi_{1,6}$	$\pi_{1,7}$	$\pi_{1,8}$	...	$\pi_{1,T}$
$\pi_{2,0}$	$\pi_{2,1}$	$\pi_{2,2}$	$\pi_{2,3}$	$\pi_{2,4}$	$\pi_{2,5}$	$\pi_{2,6}$	$\pi_{2,7}$	$\pi_{2,8}$	...	$\pi_{2,T}$
$\pi_{3,0}$	$\pi_{3,1}$	$\pi_{3,2}$	$\pi_{3,3}$	$\pi_{3,4}$	$\pi_{3,5}$	$\pi_{3,6}$	$\pi_{3,7}$	$\pi_{3,8}$	...	$\pi_{3,T}$
$\pi_{4,0}$	$\pi_{4,1}$	$\pi_{4,2}$	$\pi_{4,3}$	$\pi_{4,4}$	$\pi_{4,5}$	$\pi_{4,6}$	$\pi_{4,7}$	$\pi_{4,8}$	...	$\pi_{4,T}$
$\pi_{5,0}$	$\pi_{5,1}$	$\pi_{5,2}$	$\pi_{5,3}$	$\pi_{5,4}$	$\pi_{5,5}$	$\pi_{5,6}$	$\pi_{5,7}$	$\pi_{5,8}$	...	$\pi_{5,T}$
$\pi_{6,0}$	$\pi_{6,1}$	$\pi_{6,2}$	$\pi_{6,3}$	$\pi_{6,4}$	$\pi_{6,5}$	$\pi_{6,6}$	$\pi_{6,7}$	$\pi_{6,8}$	...	$\pi_{6,T}$
$\pi_{7,0}$	$\pi_{7,1}$	$\pi_{7,2}$	$\pi_{7,3}$	$\pi_{7,4}$	$\pi_{7,5}$	$\pi_{7,6}$	$\pi_{7,7}$	$\pi_{7,8}$	...	$\pi_{7,T}$
$\pi_{8,0}$	$\pi_{8,1}$	$\pi_{8,2}$	$\pi_{8,3}$	$\pi_{8,4}$	$\pi_{8,5}$	$\pi_{8,6}$	$\pi_{8,7}$	$\pi_{8,8}$	...	$\pi_{8,T}$
$\pi_{9,0}$	$\pi_{9,1}$	$\pi_{9,2}$	$\pi_{9,3}$	$\pi_{9,4}$	$\pi_{9,5}$	$\pi_{9,6}$	$\pi_{9,7}$	$\pi_{9,8}$	...	$\pi_{9,T}$
$\pi_{10,0}$	$\pi_{10,1}$	$\pi_{10,2}$	$\pi_{10,3}$	$\pi_{10,4}$	$\pi_{10,5}$	$\pi_{10,6}$	$\pi_{10,7}$	$\pi_{10,8}$	...	$\pi_{10,T}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\pi_{N,0}$	$\pi_{N,1}$	$\pi_{N,2}$	$\pi_{N,3}$	$\pi_{N,4}$	$\pi_{N,5}$	$\pi_{N,6}$	$\pi_{N,7}$	$\pi_{N,8}$	...	$\pi_{N,T}$

In practice, growth rates or indices are of interest, not actual prices. We therefore define the growth rate $\gamma_{j,t}$ of house j at time t as that house's price at the end of period t divided by that house's price at the end of the previous period $t-1$: $\gamma_{j,t} = \pi_{j,t}/\pi_{j,t-1}$. In similar "big matrix" fashion we may thus write out:

time									
1	2	3	4	5	6	7	8	...	T
$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{1,3}$	$\gamma_{1,4}$	$\gamma_{1,5}$	$\gamma_{1,6}$	$\gamma_{1,7}$	$\gamma_{1,8}$	$\cdots$	$\gamma_{1,T}$
$\gamma_{2,1}$	$\gamma_{2,2}$	$\gamma_{2,3}$	$\gamma_{2,4}$	$\gamma_{2,5}$	$\gamma_{2,6}$	$\gamma_{2,7}$	$\gamma_{2,8}$	$\cdots$	$\gamma_{2,T}$
$\gamma_{3,1}$	$\gamma_{3,2}$	$\gamma_{3,3}$	$\gamma_{3,4}$	$\gamma_{3,5}$	$\gamma_{3,6}$	$\gamma_{3,7}$	$\gamma_{3,8}$	$\cdots$	$\gamma_{3,T}$
$\gamma_{4,1}$	$\gamma_{4,2}$	$\gamma_{4,3}$	$\gamma_{4,4}$	$\gamma_{4,5}$	$\gamma_{4,6}$	$\gamma_{4,7}$	$\gamma_{4,8}$	$\cdots$	$\gamma_{4,T}$
$\gamma_{5,1}$	$\gamma_{5,2}$	$\gamma_{5,3}$	$\gamma_{5,4}$	$\gamma_{5,5}$	$\gamma_{5,6}$	$\gamma_{5,7}$	$\gamma_{5,8}$	$\cdots$	$\gamma_{5,T}$
$\gamma_{6,1}$	$\gamma_{6,2}$	$\gamma_{6,3}$	$\gamma_{6,4}$	$\gamma_{6,5}$	$\gamma_{6,6}$	$\gamma_{6,7}$	$\gamma_{6,8}$	$\cdots$	$\gamma_{6,T}$
$\gamma_{7,1}$	$\gamma_{7,2}$	$\gamma_{7,3}$	$\gamma_{7,4}$	$\gamma_{7,5}$	$\gamma_{7,6}$	$\gamma_{7,7}$	$\gamma_{7,8}$	$\cdots$	$\gamma_{7,T}$
$\gamma_{8,1}$	$\gamma_{8,2}$	$\gamma_{8,3}$	$\gamma_{8,4}$	$\gamma_{8,5}$	$\gamma_{8,6}$	$\gamma_{8,7}$	$\gamma_{8,8}$	$\cdots$	$\gamma_{8,T}$
$\gamma_{9,1}$	$\gamma_{9,2}$	$\gamma_{9,3}$	$\gamma_{9,4}$	$\gamma_{9,5}$	$\gamma_{9,6}$	$\gamma_{9,7}$	$\gamma_{9,8}$	$\cdots$	$\gamma_{9,T}$
$\gamma_{10,1}$	$\gamma_{10,2}$	$\gamma_{10,3}$	$\gamma_{10,4}$	$\gamma_{10,5}$	$\gamma_{10,6}$	$\gamma_{10,7}$	$\gamma_{10,8}$	$\cdots$	$\gamma_{10,T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\gamma_{N,1}$	$\gamma_{N,2}$	$\gamma_{N,3}$	$\gamma_{N,4}$	$\gamma_{N,5}$	$\gamma_{N,6}$	$\gamma_{N,7}$	$\gamma_{N,8}$	$\cdots$	$\gamma_{N,T}$

Notice that our "big matrix" representation and notation for the population diverge from the literature in an important way. For a fixed time t, previous authors have treated the growth rate as a population quantity for the entire set of houses, not for individual homes. That is, "the" growth rate has always been assumed to be a single value that described the whole set of houses; no one has ever notationally quantified the growth rate of individual houses. We believe this distinction is helpful. Explicitly acknowledging that growth rates differ across houses naturally leads to the recognition that at each time t there exists a distribution of growth rates in the population, and raises the question of the population measure for the target.

A natural target under these circumstances is some measure of the central tendency of this distribution, and perhaps the most obvious candidate is simply the arithmetic mean. That is, at each time t we may summarize the distribution of growth rates with $\bar{\gamma}_t = (1/N)\sum_{j=1}^{N}\gamma_{j,t}$. Thus, it is intuitively reasonable to define $\bar{\gamma}_t$ as a target parameter to quantify the notion of "typical house price appreciation during the tth period."[1]

A target for an index is equally natural. An index is simply a cumulative growth rate, where the accumulation is done by multiplying rather than the customary adding. Therefore, at each time t every house has its own index value, namely the product of its first t growth rates: $\iota_{j,t} = \prod_{k=1}^{t}\gamma_{j,k}$. Just as before, we could write out these various index values in a "big matrix" and for any fixed time t there exists a distribution of these index values in the population. It is reasonable to summarize these index values at t with a simple arithmetic average: $\bar{\iota}_t = (1/N)\sum_{j=1}^{N}\iota_{j,t}$. The $\bar{\iota}_t$ are an intuitive quantification of the concept of "typical house price total appreciation from some time through the tth period."[2]

[1]Note that we employ the "overbar" notation throughout this article strictly as shorthand for the standard arithmetic mean; despite its obvious visual similarity to the sample statistic $\bar{X}$, the symbol $\bar{\gamma}_t$ denotes a population parameter.

[2]The quantities $\bar{\gamma}_t$ and $\bar{\iota}_t$ are the conceptual equivalents of what previous authors have termed "the" growth rate and index respectively.

The natural question that now follows is how close alternative index methods get to these (or perhaps other) target parameters, thereby raising the issue of where an index is aiming. We define the sequence $\{G_t\}$ for $t = 1, \ldots, T$ to be the result of applying the repeat sales method (for growth rates) to the entire idealized population of $\pi_{j,t}$ values, i.e., the aim of repeat sales. We ask, for a fixed t, is G_t close to the target (say, $\bar{\gamma}_t$)? Analogously, we define the sequence $\{I_t\}$ for $t = 1, \ldots, T$ to be the result of applying the repeat sales method (for index values) to the idealized population. For a fixed t, is I_t close to the target (say, $\bar{\iota}_t$)?

2 The repeat sales method

Repeat sales methods are unique among house price index construction techniques because they are based on data that directly measure the variable of interest: house price appreciation. The key to these data is that they are observations of multiple transactions on the same property. Prices from different time periods are combined to create "matched pairs," providing a direct measure of price changes for a given property over a known period of time. Bailey et al. (1963) proposed the basic repeat sales method over three decades ago, but only after work by Case and Shiller (1987, 1989, 1990) did the idea receive significant attention in the housing research community.

2.1 Where repeat sales is aiming

To determine whether repeat sales indices are aiming at the desired target, we must first resolve where they are aiming (i.e. the I_t and G_t). Using our notation, the model underlying the Bailey et al. method can be written as follows. Letting t_1 and t_2 be the times of the first and second transactions, a price relative $\pi_{j,t_2}/\pi_{j,t_1}$ can be modeled as

$$\frac{\pi_{j,t_2}}{\pi_{j,t_1}} = \frac{I_{t_2}}{I_{t_1}} \times u_{j,t_1,t_2}$$

where u_{j,t_1,t_2} is an idiosyncratic error term. Taking logs,

$$\log\left(\frac{\pi_{j,t_2}}{\pi_{j,t_1}}\right) = -\log(I_{t_1}) + \log(I_{t_2}) + \log(u_{j,t_1,t_2}).$$

Using vector notation, this relationship can be expressed as $y = X\beta + \varepsilon$ where y is the (known) vector of logged price relatives; each row of X is of dimension T such that the tth component of the row is -1 if $t = t_1$, $+1$ if $t = t_2$, and 0 otherwise; the tth component of the (to this point unknown) β vector is $\beta_t = \log(I_t)$; and ε is the (to this point unknown) vector of $\log(u_{j,t_1,t_2})$ values.

An alternate specification can be used for growth rates instead of indices, modeling a logged price relative as

$$\log\left(\frac{\pi_{j,t_2}}{\pi_{j,t_1}}\right) = \sum_{t=t_1+1}^{t_2} \log(G_t) + \log(u_{j,t_1,t_2}).$$

Again this can be written in vector notation, this time as $y = Z\Gamma + \varepsilon$ where the vectors y and ε are as above; the rows of Z are such that the tth component is $+1$ if and only if $t_1 < t \leq t_2$ and 0 otherwise; and the tth component of the (unknown) Γ vector is $\Gamma_t = \log(G_t)$.

Bailey et al. used ordinary least squares to define the coefficient vector β of logged indices. Recall that with ideal data the price of every house is observed in every time period, so that $t_2 = t_1 + 1$ always and the price relatives are all just the $\gamma_{j,t}$. This results in a very special structure for the X matrix. It turns out to be straightforward (though messy) to show in closed form that the solution to the usual normal equations $\beta = \log(I) = (X'X)^{-1}X'y$ consists of period-by-period means of the logged index values; i.e.,

$$\beta_t = \log(I_t) = \frac{1}{N}\sum_{j=1}^{N} \log(\iota_{j,t}) = \overline{\log(\iota_t)}.$$

The repeat sales house price index is then defined by componentwise exponentiation of β. Therefore, the Bailey et al. method defines the value of I_t to be $\exp(\beta_t) = \exp(\overline{\log(\iota_t)})$.

Growth rates are similar. The Z matrix has very special structure; the coefficient vector $\Gamma = \log(G) = (Z'Z)^{-1}Z'y$ can be determined in closed form; and it turns out to comprise period-by-period means of the logged growth rates; i.e.,

$$\Gamma_t = \log(G_t) = \frac{1}{N}\sum_{j=1}^{N} \log(\gamma_{j,t}) = \overline{\log(\gamma_t)}.$$

Again exponentiating componentwise, repeat sales defines the value of G_t to be $\exp(\Gamma_t) = \exp(\overline{\log(\gamma_t)})$. Note that I_t is close but not exactly equal to the possible target value $\bar{\iota}_t$, and that G_t is close but not exactly equal to the possible target value $\bar{\gamma}_t$. In Section 3 we elaborate on this discrepancy.

2.2 The sample data

With real data we have nothing at all like the idealized population described in Section 1. First, the population of houses changes over time as new houses are constructed and old houses are demolished, resulting in a changing housing stock. Second, there is inevitably a process of "sampling," or what may be thought of as a stage-by-stage "removal of information" compared to the population.

The existence of construction and demolition means that the total number of houses in period t is variable and depends on t. Thus, at time t, the population consists of houses indexed by $j = 1, \ldots, N_t$. Preliminarily assuming that we knew every house's price at every time period, an illustrative example of the "big matrix" may be written as follows:

					time					
0	1	2	3	4	5	6	7	8	…	T
$\pi_{1,0}$	$\pi_{1,1}$	$\pi_{1,2}$	$\pi_{1,3}$	$\pi_{1,4}$	$\pi_{1,5}$	$\pi_{1,6}$	$\pi_{1,7}$			
$\pi_{2,0}$	$\pi_{2,1}$	$\pi_{2,2}$	$\pi_{2,3}$	$\pi_{2,4}$	$\pi_{2,5}$	$\pi_{2,6}$	$\pi_{2,7}$	$\pi_{2,8}$	…	$\pi_{2,T}$
	$\pi_{3,1}$	$\pi_{3,2}$	$\pi_{3,3}$	$\pi_{3,4}$	$\pi_{3,5}$					
	$\pi_{4,1}$	$\pi_{4,2}$	$\pi_{4,3}$	$\pi_{4,4}$	$\pi_{4,5}$	$\pi_{4,6}$	$\pi_{4,7}$	$\pi_{4,8}$	…	$\pi_{4,T}$
		$\pi_{5,2}$	$\pi_{5,3}$	$\pi_{5,4}$	$\pi_{5,5}$	$\pi_{5,6}$	$\pi_{5,7}$	$\pi_{5,8}$		
		$\pi_{6,2}$	$\pi_{6,3}$	$\pi_{6,4}$	$\pi_{6,5}$	$\pi_{6,6}$	$\pi_{6,7}$	$\pi_{6,8}$	…	$\pi_{6,T}$
			$\pi_{7,3}$	$\pi_{7,4}$	$\pi_{7,5}$	$\pi_{7,6}$				
			$\pi_{8,3}$	$\pi_{8,4}$	$\pi_{8,5}$	$\pi_{8,6}$	$\pi_{8,7}$	$\pi_{8,8}$	…	$\pi_{8,T}$
				$\pi_{9,4}$	$\pi_{9,5}$	$\pi_{9,6}$	$\pi_{9,7}$	$\pi_{9,8}$		
				$\pi_{10,4}$	$\pi_{10,5}$	$\pi_{10,6}$	$\pi_{10,7}$	$\pi_{10,8}$	…	$\pi_{10,T}$
				⋮	⋮	⋮	⋮	⋮	⋮	⋮

The various rows in this matrix represent canonical cases. For example, house 2 represents all houses that existed during the base period and were demolished during the 8th period; house 7 represents all houses that were constructed during the 3rd period and demolished during the 7th; and so on. Also note that if no other houses existed at any time, some sample values for N_t in this example would be $N_1 = 4$, $N_3 = 8$, $N_4 = 10$, $N_8 = 7$, $N_T = 5$.

The "sampling" process in real data reduces information further. The first stage loss of information comes from the fact that houses are sold infrequently, so the price of a given house remains unobserved in most periods. This reduced information might appear as follows:

					time					
0	1	2	3	4	5	6	7	8	…	T
		$\pi_{1,2}$			$\pi_{1,5}$		$\pi_{1,7}$			
$\pi_{2,0}$				$\pi_{2,4}$		$\pi_{2,6}$				
	$\pi_{3,1}$									
			$\pi_{4,3}$			$\pi_{4,6}$		$\pi_{4,8}$		
				$\pi_{5,4}$		$\pi_{5,6}$				
			$\pi_{6,3}$	$\pi_{6,4}$						$\pi_{6,T}$
					$\pi_{7,5}$					
			$\pi_{8,3}$					$\pi_{8,8}$		
						$\pi_{9,6}$		$\pi_{9,8}$		
				$\pi_{10,4}$			$\pi_{10,7}$			
				⋮	⋮	⋮	⋮	⋮	⋮	⋮

The next stage loss of information is the loss of certain houses (because they are not included in the sample), and the loss of price observations for some houses that are included. Suppose, for example, that houses 2, 4, 5, 7 and 9 are not included in the sample, and that the 5th period price for house 1 and the 3rd period price for house 8 are also unobserved. Then the reduced information appears as:

time										
0	1	2	3	4	5	6	7	8	$\dots$	T
		$\pi_{1,2}$					$\pi_{1,7}$			
	$\pi_{3,1}$									
			$\pi_{6,3}$	$\pi_{6,4}$						$\pi_{6,T}$
								$\pi_{8,8}$		
				$\pi_{10,4}$			$\pi_{10,7}$			
				$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

It is worth stressing that to this point the loss of information incurred by sampling is not unique to repeat sales. That is, every other house price index method also suffers the same losses. But the repeat sales method does impose one unique additional loss of information. Because it relies on pairs of prices on the same property to create price relatives, it "throws out" properties for which there is only one price observation. For example, in this illustration it is possible to create only four price relatives: $\pi_{1,7}/\pi_{1,2}$, $\pi_{6,4}/\pi_{6,3}$, $\pi_{6,T}/\pi_{6,4}$ and $\pi_{10,7}/\pi_{10,4}$. The price observations on houses 3 and 8 cannot be included in the repeat sales estimation, whereas with some other methods these prices could be used.

Finally, notice also that the vast majority of the observations in the data actually used to create the indices yield direct estimates of neither a growth rate nor of an index. This is because it is typically not the case that the repeated observations are exactly one period apart, nor is it typical that the first observation happens to be in the base period. Nevertheless it is still possible to view the repeat sales method as producing estimates that are simply means. The two twists are that these means are now weighted and the quantities entering into the means are not simply period-by-period growth rates, but are something close.

2.3 What repeat sales yields with sample data

The simplest way of illustrating what repeat sales actually does is through a concrete demonstration. Consider a "toy" example in which there are only two time periods, so we wish to produce estimates of only two growth rates (indices). We then have two observation times 1 and 2, plus the base period time 0. Let S_{01} refer to the set of all homes that were observed at times 0 and 1; let S_{02} refer to the set of all homes that were observed at times 0 and 2; and let S_{12} refer to the set of all homes that were observed at times 1 and 2. (Notice that a house that was observed at all three times would appear in both S_{01} and S_{12}, but not in S_{02}.) Denote the cardinalities of these sets by n_{01}, n_{02}, and n_{12} respectively. Finally, let $\bar{y}_{ij}$ be the arithmetic mean log price relative of all homes in set $S_{ij} : \bar{y}_{ij} = (1/n_{ij}) \sum_{k \in S_{ij}} \log(\pi_{k,t_j}/\pi_{k,t_i})$. Clearly, in this case we have three such $\bar{y}_{ij}$: $\{\bar{y}_{01}, \bar{y}_{02}, \bar{y}_{12}\}$.

Although it is not obvious, Bailey et al. showed that the estimates of the two mean logged price indices are simply weighted averages of these three $\bar{y}_{ij}$'s. That is, even though their method requires a least squares regression solution, it turns out that these solutions are in fact nothing more than complicated weighted averages of the three average logged price relatives: the two components of the solution vector $\widehat{\log(I)}$ can be written as

$$\begin{aligned}
\widehat{\log(I_1)} &= \frac{n_{01}(n_{02}+n_{12})}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}\bar{y}_{01} + \frac{n_{02}n_{12}}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}(\bar{y}_{02}-\bar{y}_{12}) \\
\widehat{\log(I_2)} &= \frac{n_{02}(n_{01}+n_{12})}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}\bar{y}_{02} + \frac{n_{01}n_{12}}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}(\bar{y}_{01}+\bar{y}_{12})
\end{aligned}$$

Moreover, these formulas make intuitive sense. For example, in this form we see that $\widehat{\log(I_1)}$ is nothing more than the weighted average of the two quantities $\bar{y}_{01}$ and $\bar{y}_{02}-\bar{y}_{12}$. The obvious estimator of $\log(I_1)$ is $\bar{y}_{01}$, the average logged appreciation amount of houses that were observed in period 1. Another, less obvious estimator of $\log(I_1)$ is $\bar{y}_{02}-\bar{y}_{12}$, since this amounts to taking the typical logged total appreciation in those houses that were not observed until period 2, then adjusting that estimate by subtracting off an estimate of the typical logged appreciation amount over period 2 only.

These two estimators of $\log(I_1)$ are combined in an intuitively pleasing way: the weights given to the two estimators are essentially proportional to their respective sample sizes. This accords with the well-known statistical fact that the minimum variance weighted average of two uncorrelated estimators is the one in which each estimator is weighted in inverse proportion to its variance, or equivalently, in direct proportion to its sample size. As an illustration, if each set of homes in our toy example had identical sample size (i.e. $n_{01} = n_{12} = n_{02}$) then the weight on $\bar{y}_{01}$ would be

$$\frac{n_{01}(n_{02}+n_{12})}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}} = \frac{2n_{01}^2}{3n_{01}^2} = \frac{2}{3}$$

and the weight on $\bar{y}_{02}-\bar{y}_{12} = 1/3$.

Note that $\widehat{\log(I_2)}$ is expressed as a similar formula, and similar intuitive reasoning holds: it can be thought of as a weighted average of two intuitively sensible estimates of $\log(I_2)$ defined in terms of $\bar{y}_{01}$, $\bar{y}_{12}$, and $\bar{y}_{02}$.

It should come as no surprise that similarly, although Bailey et al. did not show it, the estimates of the two logged growth rates are also just weighted averages of two terms involving the three $\bar{y}_{ij}$: $\widehat{\log(G_1)}$ is the same as $\widehat{\log(I_1)}$ above (the first period growth rate is by definition the same as the first period index), while

$$\widehat{\log(G_2)} = \frac{n_{12}(n_{01}+n_{02})}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}\bar{y}_{12} + \frac{n_{01}n_{02}}{n_{01}n_{02}+n_{01}n_{12}+n_{02}n_{12}}(\bar{y}_{02}-\bar{y}_{01}).$$

This example can be simply extended to three time periods. In this case we have four observation times and six different S_{ij}: S_{01}, S_{02}, S_{03}, S_{12}, S_{13}, S_{23}, each with a corresponding

$\bar{y}_{ij}$. It is straightforward but tedious to show that each of the three logged growth rate estimates can again be expressed as weighted averages of terms involving the various $\bar{y}_{ij}$, where in each case the same pleasing intuition holds.

Although it has not been demonstrated, the obvious conjecture is that for T time intervals the logged growth rate estimates will be a complicated weighted average of the $\binom{T+1}{2}$ different $\bar{y}_{ij}$. Our interest is not in the precise form of this general relationship. Rather, it is recognizing that when applied to real data the Bailey et al. repeat sales procedure provides intuitively reasonable estimates of $\log(I_t)$ and $\log(G_t)$, or I_t and G_t after exponentiating.

3 Choosing a target

We turn now to concerns raised in the literature about repeat sales procedures, focusing first on those that revolve implicitly around the choice of a target. Differentiating among targets is not an inconsequential exercise: if an index does not measure the desired quantity it will almost certainly be less effective in its desired application. Thus as a pragmatic matter, choosing among competing targets is of critical importance to practitioners.

We are aware of no explicit discussions in the literature regarding the desired target for any price index, repeat sales or otherwise. This lacuna is, in part, our justification for writing this paper. However, consistent with our discussion of the idealized population data in Section 1, there is clearly an implicit view that the appropriate target for repeat sales procedures is either the arithmetic mean growth rate $\bar{\gamma}_t$ or the arithmetic mean index value $\bar{\iota}_t$. Of course no single, universally appropriate target for price indices exists; that is one of the points of this section. Nonetheless, the targets $\bar{\gamma}_t$ and $\bar{\iota}_t$ provide a valuable point of reference in comparing alternative methods.

3.1 Geometric versus arithmetic means

As previously noted, the Bailey et al. repeat sales procedure applied to the idealized population yields $G_t = \exp(\overline{\log(\gamma_t)})$ and $I_t = \exp(\overline{\log(\iota_t)})$, rather than $\bar{\gamma}_t$ and $\bar{\iota}_t$. This discrepancy can be precisely stated: for fixed t, G_t and I_t are in fact the *geometric* rather than arithmetic means of the raw growth rates $\{\gamma_{j,t}\}$ and index values $\{\iota_{j,t}\}$.

To see this, note that the geometric mean of a set of positive numbers $x_1, x_2, \ldots, x_n$ can be expressed as

$$\begin{aligned}
\tilde{x} &= \left(\prod_{i=1}^{n} x_i\right)^{1/n} \\
&= \exp\left(\log\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right) \\
&= \exp\left(\frac{1}{n}\log\left(\prod_{i=1}^{n} x_i\right)\right)
\end{aligned}$$

$$= \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right).$$

This is exactly the form of the G_t and I_t obtained from the repeat sales procedure applied to the idealized population. For this reason Shiller (1991) termed the Bailey et al. method "geometric repeat sales." To reiterate: applying repeat sales to the idealized population yields period-by-period arithmetic means, but of the *logged* growth rates (and exponentiating, these are precisely the period-by-period geometric mean growth rates). Thus, with sample data the most natural characterization is that repeat sales provides an *estimate* of the population *geometric* mean growth rate. We can express this concisely: when measuring growth rates repeat sales delivers $\hat{G}_t$, an estimate of $G_t = \tilde{\gamma}_t$ not $\bar{\gamma}_t$; when measuring index values repeat sales delivers $\hat{I}_t$, an estimate of $I_t = \tilde{\iota}_t$ not $\bar{\iota}_t$.

Why does this matter? It can easily be shown using Jensen's inequality that the geometric mean of any set of positive numbers is always less than or equal to the arithmetic mean of those numbers, with equality occuring only if all the numbers are equal. Here, the values that the repeat sales method are trying to hit ($G_t = \tilde{\gamma}_t$ and $I_t = \tilde{\iota}_t$) are strictly less than the values $\bar{\gamma}_t$ and $\bar{\iota}_t$.

This recognition has explicitly raised the question of how to use repeat sales to get an arithmetic mean.[3] Two approaches have been suggested. Shiller (1991) developed an instrumental variables technique for estimating arithmetic means directly. Although this process is somewhat more complicated than the least squares procedure of Bailey et al. it is entirely workable. Nonetheless it has not been widely adopted.

Goetzmann (1992) proposed an *ex post* correction to the Bailey et al. model. Although the specific formulation he proposed is flawed, it suggests a relatively simple scalar adjustment to the geometric mean. To illustrate, suppose $X_1, X_2, \ldots, X_n$ are independently and identically lognormal(μ, σ^2), so that the $Y_i \equiv \log(X_i)$ are independently and identically normal(μ, σ^2). From any standard textbook, $E(X_i) = \exp(\mu + \sigma^2/2)$. Therefore, the expected value of the arithmetic mean of the Xs is $E(\bar{X}) = E(X_i) = \exp(\mu + \sigma^2/2)$.

On the other hand, the geometric mean of the Xs can be written $\exp((1/n)\sum_{i=1}^{n}\log(X_i)) = \exp((1/n)\sum_{i=1}^{n} Y_i) = \exp(\bar{Y})$. Since the $Y_i = \log(X_i)$ are normal(μ, σ^2), $\bar{Y}$ is normal$(\mu, \sigma^2/n)$, so that $\exp(\bar{Y})$ is lognormal$(\mu, \sigma^2/n)$. Thus, the expected value of the geometric mean of the Xs is $E(\exp(\bar{Y})) = \exp(\mu + \sigma^2/2n)$.

Comparing these two, we see that on average the arithmetic mean is larger than the geometric mean by a factor of

$$\frac{\exp(\mu + \sigma^2/2)}{\exp(\mu + \sigma^2/2n)} = \exp\left(\frac{(n-1)\sigma^2}{2n}\right),$$

[3] In a longitudinal (time series) context the geometric mean is generally a more natural measure of center, whereas in a cross-sectional context the arithmetic mean is more appropriate. In the case of housing data both contexts are present. However repeat sales indices are most often used as a measure of house price change at each time period, so the cross-sectional context is the one of interest and the focus on arithmetic means can be justified.

and that this factor approaches $\exp(\sigma^2/2)$ as $n \to \infty$. Consequently, a simple correction for obtaining an estimate of the arithmetic mean is to multiply the estimated geometric mean by the above factor.

3.2 Measures of price: appraisal versus purchase

Any repeat sales price index is based on house "prices," but in practice price observations come in two forms: purchase and appraised. This has raised the question of whether the purchase price, the appraisal price or both are the appropriate target. For example, Chinloy et al. (1997) offered evidence to suggest that appraisal prices differ systematically from purchase prices. In particular, they found that appraisals tend to be higher than purchase prices, possibly because incentives exist to support a transaction taking place in refinance situations. They also found that in contrast to conventional wisdom, appraisal data are not significantly less volatile than purchase data. In their minds this raised questions about the use of appraisal data, which they formulated by suggesting that price indices based on appraisals are "biased" relative to those obtained from using purchase prices. For this reason the authors proposed that either alternative estimation procedures be used, or that indices be estimated using only purchase prices. But clearly, the authors' implicit target was purchase price. Practitioners interested in either appraisals or both purchase and appraisal prices will have a different perspective on the "bias" identified by Chinloy et al.[4]

3.3 Nontemporal components of appreciation

Goetzman and Spiegel (1995) posited that house price appreciation is made up of both temporal and nontemporal components. They contended that home improvements are a major nontemporal (fixed) component of house price growth, and that improvements are most likely to occur either immediately before or after a sale. Implicit in the authors' view is that the appropriate target for a price index is only the temporal component of price appreciation, and they argued that the inclusion of nontemporal components leads to a positive "bias" away from this target. As a solution to this problem they proposed the inclusion of a constant (intercept) term in the Bailey et al. estimation or the use of maximum likelihood methods. Once more, however, the validity of this concern depends on the target of interest.

3.4 Holding quality constant

The issue of home improvement raises the more general concern of "holding quality constant." Prices indices have many uses, and the degree to which holding quality constant is desirable depends on the application. For example, some researchers use price indices, either implicitly

[4]On a related note, Clapp and Giaccotto (1992) proposed what they termed the "assessed value" methodology as an "alternative" to repeat sales. Their idea was that an assessed value (from an appraiser) could be used to create matched pairs for those properties for which there was only one observation on price, thereby increasing the repeat transactions sample size tremendously. In spirit this expands the target to incorporate more measures of house value, as opposed to Chinloy et al. who explicitly narrowed the definition of the target.

or explicitly, as proxies of the price per unit of housing stock in studies that model housing as a homogeneous commodity. In such instances researchers typically wish to compare the prices of identical houses either cross-sectionally or over time. Alternatively, researchers use price indices to estimate the changing value of a "representative" house or a portfolio of houses over time. In this case researchers may wish to allow for changes in quality through improvements or depreciation that are typical for the portfolio.

Broadly speaking, it is possible to hold quality constant through two techniques: explicit conditioning on the sample or controlling for different housing characteristics. Conditioning can be accomplished by calculating an index from a sample that is, for example, limited to single-family detached houses of between 1,500 and 2,500 square feet that have been constructed in the past 10 years. This technique is of course applicable to repeat sales methods. However repeat sales procedures also directly control for different housing characteristics through the use of paired sampling. By observing the same house at two points in time, all housing characteristics affecting house price growth (other than improvements and depreciation) are held constant.

Much has been made of the "inability" of repeat sales procedures to hold constant for improvements and depreciation. Several authors have argued that hedonic methods are preferable for this reason because these techniques explicitly control for the relationship between housing characteristics and price (see more discusion on this issue in Section 5). Case and Quigley (1991) extended this belief to its logical conclusion by, among other things, explicitly incorporating hedonic-like controls into repeat sales methods.

Again, this debate is implicitly over the appropriate target for an index. Researchers looking for a proxy of the per unit price of housing may appreciate the hedonic methodology's theoretically greater ability to control for quality. Researchers looking for an estimate of the change in the value of housing may prefer to include the impact of improvements and depreciation in their indices.

4 Problems with achieving the aim

We are now in position to address the question how well repeat sales estimators approximate their aims. This turns the focus to measures of statistical accuracy such as bias, consistency and efficiency. Much has been written about these subjects. Our contribution is in framing these issues, and in specifying a precise metric for comparison. Specifically, we ask how well $\hat{G}_t$ and $\hat{I}_t$ estimate G_t and I_t. We also note that unlike the problems that are actually due to the choice of target as mentioned in Section 3, these issues are inherent to all repeat sales methods. That is, the "biases" previously described in Section 3 have the potential of being considered irrelevant or at least not of concern, by defining the desired target appropriately. In contrast, the potential statistical problems discussed in Section 4 are always present; they cannot be defined away.

4.1 Sample selection bias

Problems of sample selection are of concern because nonprobability sampling introduces the possibility of bias. We use the term "bias" in the strict statistical sense: $\hat{\theta}$ is a *biased* estimator of θ if the expected value of $\hat{\theta}$ is not equal to θ. In this case, the concern is if $E(\hat{G}_t) \neq G_t$ or if $E(\hat{I}_t) \neq I_t$ due to sample selection issues.

As previously noted, there are several dimensions of "sampling" in the creation of repeat sales data. The first stage of sampling is based on whether or not a house price is observable in a given time period. Only houses with observable prices have any potential of being in the ultimately realized sample. The second stage is sampling among those houses for which prices are actually observed. Finally, a third stage introduces additional sampling as properties for which there is only one price observation are thrown out.

The empirical evidence on sample selection bias is in its infancy and it is not yet clear its extent, or which stage of sampling is the most problematic. The issue of sample selection bias in the context of price indices was first raised significantly by Gatzlaff and Haurin (1994). These authors focused on the first stage of sampling and posited that the probability of observing a house's price varied with the price appreciation it experienced, and that the resulting sample was therefore significantly nonrandom. They suggested a censored regression procedure to account for this problem. They first applied their framework to hedonic regression techniques, but in Gatzlaff and Haurin (1997) extended it to repeat sales procedures. Meese and Wallace (1997) addressed problems of sampling in the third stage, focusing on issues unique to repeat sales.

We provide a brief example to illustrate the possible effects of sample selection. This "toy" example focuses on the first stage of sampling because it is the easiest to explain and the issues in each stage are conceptually similar.

Suppose there are only two types of houses whose prices are observed at three points in time. The values given in the table below are prices in thousands of dollars for this population.

	time		
	0	1	2
Type 1 houses	100	102	107.1
Type 2 houses	100	101	104.03

The first type of house experiences rapid appreciation rates: each appreciate by 2% over period 1 and 5% over period 2, so the $\gamma_{j,1}$ and $\gamma_{j,2}$ values for these houses are all 1.02 and 1.05. The second type of house has slower appreciation rates: each appreciate only 1% over period 1 and 3% over period 2, so the $\gamma_{j,1}$ and $\gamma_{j,2}$ values for these houses are all 1.01 and 1.03.

Letting M_1 stand for the total number of type 1 houses and M_2 for the total number of type

2 houses, we can use the formulas from Section 2: for period 1,

$$G_1 = \exp(\overline{\log(\gamma_1)}) = \exp\left(\frac{1}{M_1 + M_2}\sum_{j=1}^{M_1+M_2}\log(\gamma_{j,1})\right).$$

In the summation, there are M_1 occurrences of log(1.02) and M_2 occurrences of log(1.01). Hence $\log(G_1) = (\log(1.02)M_1 + \log(1.01)M_2)/(M_1 + M_2)$; and similarly for period 2, $\log(G_2) = (\log(1.05)M_1 + \log(1.03)M_2)/(M_1 + M_2)$. In other words, $\log(G_1)$ and $\log(G_2)$ would simply be weighted averages of the logged price relatives of the two types of houses, where the weights were proportional to the respective population sizes. To further simplify this example, suppose that the two types of houses were equally numerous ($M_1 = M_2$). Then $\log(G_1)$ and $\log(G_2)$ are simply the unweighted means: $\log(G_1) = (\log(1.02) + \log(1.01))/2 = .01488$, so that $G_1 = \exp(.01488) = 1.01499$, and $\log(G_2) = (\log(1.05) + \log(1.03))/2 = .03917$, so that $G_2 = \exp(.03917) = 1.03995$.

In the sample it will never be the case that the price of every house will be observed in every time period. During period t, only some fraction $p_{1,t}$ of the type 1 houses will be observed, and similarly some fraction $p_{2,t}$ of the type 2 houses will be observed. For the moment assume that $p_{1,1} = p_{1,2} = 1$ and that $p_{2,1} = 0$ while $p_{2,2} = 1$: during period 1 all of the faster appreciating houses are observed but none of the slower appreciating ones, while in period 2 all of both types of houses are observed.[5]

The problem of sample selection bias can be explicitly illustrated by noting that at time t_1 the only information available is that from type 1 houses because none of the type 2 houses have been sold yet. As before, let $\bar{y}_{ij}$ be the mean of the logged price relatives for all houses with initial transaction at t_i and final transactions at t_j, and n_{ij} be the corresponding number of houses. In this case the repeat sales estimate $\widehat{\log(G_1)}$ reduces to simply $\bar{y}_{01} = \log(1.02)$. Therefore, the estimate is just $\hat{G}_1 = \exp(\log(1.02)) = 1.02$. Now the sample selection bias is evident: at time t_1, since no other "sampling" is taking place, the expected value of the repeat sales estimate $\hat{G}_1$ is $E(\hat{G}_1) = 1.02$ while the true value of G_1 is 1.01499.[6] Thus $\hat{G}_1$ is biased.

The source of the bias is that $p_{1,1} \neq p_{2,1}$. Because the two types of houses have unequal probabilities of being observed, the estimate of the population appreciation G_1 is obviously weighted more heavily toward the characteristics of the houses that are observed. If the appreciation of the two types of houses differs, the estimate will be biased.

[5]We have intentionally chosen these proportions to reflect a simplified version of a conjecture that seems intuitively reasonable, but for which there does not appear to be any clear empirical evidence: houses that appreciate at a faster rate tend to be observed more often. A more realistic assumption might be that $p_{1,t} > p_{2,t} > 0$, or even more generally, simply that $p_{1,t} \neq p_{2,t}$.

[6]This example generalizes directly to other methodologies for computing price indices. Faced with this type of sampling, virtually any "reasonable" technique for estimating G_1 will yield an expected value of 1.02 instead of the correct value of 1.01499.

4.2 Revision volatility

An additional issue that has arisen in the literature is the so-called "revision volatility" problem. As time advances and new information is added as the result of more houses entering the repeat sales database, previously computed indices change. In the absence of sample selection issues such revisions to past indices do not present a theoretical problem so much as a practical one. That is, it is easy to understand why an index based on sample information changes as a result of new observations: it simply reflects the fact that estimates change when the sample changes. As noted by Shiller (1993) such revisions are common to many indices, and should in fact be welcomed because they are the result of increased efficiency in the estimators.

However in some applications the fact that prices are retroactively volatile may create practical problems. Even more to the point, as noted by Clapp and Giaccotto (1996), the presence of nonprobability sampling can exacerbate these difficulties. In particular, as illustrated in our previous example, the existence of any sample selection bias means that the initial (unrevised) estimate will be biased. Their proposed solution to this problem is, in its most general form, equivalent to a weighted least squares regression procedure that downweights observations from housing types that are sampled "too frequently."

The issue of revision volatility can be easily illustrated using our previous example: at time t_1 the estimated value of G_1 was 1.02 (and this was incorrect). But at time t_2, the slower appreciating type 2 houses have also been sold, and these houses provide information on period 1 appreciation. When we perform the repeat sales regression, the estimate of G_1 is now revised to 1.01016, while the estimate of G_2 is 1.03987.

In light of the formulas in Section 2 it is clear why these revisions take place. After one time interval $\widehat{\log(G_1)}$, the estimate of $\log(G_1)$, was simply $\bar{y}_{01}$; but after two intervals the estimate of $\log(G_1)$ became a weighted average of the original $\bar{y}_{01}$ and the new information present in $\bar{y}_{02} - \bar{y}_{12}$. In general (though not in this particular example), as we continue to collect data each additional interval will add further information in the form of additional $\bar{y}_{ij}$ that enter into the weighted average estimate of $\log(G_t)$.

4.3 Asymptotic inconsistency

We now raise an issue that might be called "asymptotic inconsistency," one that, as far as we know, has never been addressed in the literature. By asymptotic we mean as time goes to infinity. Note we do not mean that the sample size must approach the population size in every period, since that is precisely equivalent to requiring that the price of every house be observed in every period. Rather, the question is this: if we make the assumption that the price of every house that was existent at a particular time is *eventually* observed for a second time and therefore appears in the repeat sales database, is it true that $\hat{G}_t$ converges to G_t? We conjecture that the answer is "no" in the presence of nonprobability sampling.

One of the common technical definitions of consistency requires that an estimator meet two conditions, one of which is being asymptotically unbiased. We conjecture that the repeat sales estimator will be biased unless the probability of being observed is the same for all types of houses in all time periods (i.e. $p_{i,t} = p_{j,t}$ for all i, j and t). If these probabilities are not equal in each time period (almost assuredly the case in reality), then every period's estimate may be biased, each possibly by a different amount. Even asymptotically this bias will not vanish; therefore consistency is impossible.

Our previous example clearly illustrates this circumstance. Because every house in the population was observed at time t_2, no further period price information will change the estimates of G_1 or G_2; that is, no further revision will take place in this particular example.[7] Therefore, the estimate $\hat{G}_1$ (and hence its expected value, again because no other sampling is present) will remain 1.01016 whereas the true value is $G_1 = 1.01499$; similarly, the estimate $\hat{G}_2$ (and its expected value) will remain 1.03987 whereas the true value is $G_2 = 1.03995$. Hence, even "asymptotically" the estimates are biased.

At a theoretical level it is difficult to assess the significance of this problem, and a pragmatic answer awaits further empirical research on the impact of sample selection. But as a final point it is worth noting that this problem is not unique to repeat sales procedures: all methods for computing house price indices are at risk of asymptotic inconsistency in the presence of nonprobability sampling.

4.4 Heteroscedasticity

Case and Shiller (1987) first raised the issue of possible heteroscedasticity in the repeat sales estimation. Heteroscedasticity is a concern with any estimation procedure because it raises the issue of efficiency (i.e. the estimators $\hat{G}_t$ and $\hat{I}_t$ have "too large" a variance). But heteroscedasticity may not be a significant concern in all circumstances because even in the face of inefficiency the repeat sales index can be unbiased and consistent. Many applications rely solely on the indices themselves and are based on plentiful data. In these instances inefficient estimators are unlikely to create significant practical problems, so ignoring the question of heteroscedasticity may be justified. However, heteroscedasticity is certainly an issue when estimations are performed on small samples or when variances are used either for calculating confidence intervals or probabilities.

Case and Shiller's primary motivation for addressing the problem of heteroscedasticity came from their focus on the efficiency of the housing market and their desire to test whether house prices follow a random walk. A random walk model implies that the variance of house prices (and therefore growth rates) increases linearly with time. To test this assumption Case and Shiller estimated a least squares fit to the squared residuals from their initial repeat sales regression on a linear function of the holding period Δt (the number of periods between the two sales of a given house). Because they found a statistically significant relationship

[7] This general result occurs when there is any time period over which no transactions span (i.e. there exists some time period t^* such that t^* is not strictly between t_1 and t_2 for any transaction pair in the sample). In such an instance, the resulting $X'X$ matrix can be written as a block diagonal, with observations having $t_2 \leq t^*$ making up one block and observations having $t_1 > t^*$ making up the other. Under these circumstances it is easy to see that any observations with $t_1 > t^*$ will have no impact on the estimates of G_t where $t \leq t^*$.

between variance and holding period in this second stage, they also suggested a third stage to the estimation procedure: a weighted least squares regression of the original data, where the weights were the estimated standard deviations given by the results of the second stage.

This procedure appropriately corrects for any heteroscedasticity in the data provided that the model of the variance is correct. But this is a significant proviso. The empirical evidence fairly strongly rejects Case and Shiller's assumption of a random walk model; for example, see Hill et al. (1997). Nor does Case and Shiller's second stage regression explain much of the variance in the squared residuals. In response, some researchers have used alternative specifications for the second stage estimations, adding, for example, a $(\Delta t)^2$ term to the second stage regression. Goodman and Thibodeau (1995) explored a more general specification, adding additional information to explain the heteroscedastic relationship.

5 Competing methodologies

In this section we briefly review proposed alternatives to repeat sales. The various methods for constructing house price indices can be roughly divided into four categories: "summary," hedonic, repeat sales, and hybrid. Each of these four classes has its strengths and weaknesses. Each also has its strong proponents. Existing comparisons of the alternative methodologies are generally couched in terms of "bias" and efficiency, or made against some imprecisely stated benchmark. But in our view, much of the disagreement over the preferred procedure arises because the targets and aims of the underlying estimation methods have never been precisely or explicitly established.

We provide only a brief overview of the other competing techniques, along with a similar level discussion of repeat sales. As throughout this article, we focus on the aims of the methods, as well as their implicit targets. In doing so we demonstrate that legitimate rationales exist for the use of each.

5.1 "Summary" methods

Probably the simplest and most obvious way to construct an index of house prices is as follows. At some time, for a selected set of homes, compute some summary measure of central tendency of the prices; this gives an estimate of the price of a "typical" house. At all subsequently desired times, repeat this process with another, probably different sample of houses. The resulting sequence of mean or median prices then forms a price series. From the price series it is straightforward to derive a price index series (divide each period's price by the base period's price) or a growth rate series (divide each period's price by the previous period's price).[8]

The aim of any summary method is simply a measure of the central tendency of house prices for the entire population. If we again use the idealized data of Section 1 and the arithmetic

[8]Although the name "summary method" to our knowledge has never before been used in the literature, we feel it is appropriate because it describes how the underlying price series is constructed.

mean as the desired measure, for example, then all summary methods are trying to deliver $\bar{\pi}_t$ initially. From this, the growth rate over period t can be defined as $\bar{\pi}_t/\bar{\pi}_{t-1}$. Similarly, the index at period t can be defined by $\bar{\pi}_t/\bar{\pi}_0$.

Probably the single most widely used (summary) measure of house price trends is the one published by the National Association of Realtors. It is a median sales price series, and uses data from sales of all existing homes. Because summary methods are so straightforward from a theoretical viewpoint, no articles that we are aware of have been devoted to their study. However, summary methods have been mentioned by some authors as points of comparison; for examples, see Mark and Goldberg (1984), Hosios and Pesando (1991), Crone and Voith (1992), and Gatzlaff and Ling (1994).

The most obvious strength of an index computed by a summary method is its directness and ease of interpretability: the underlying price series is simply some easily understood measure of central tendency computed at various points in time. Thus the aim and the target are in close alignment if the desired target is this same simple measure.[9]

However, recognizing that in practice a different sample will generally be used at each time, this simplicity is also this method's main weakness. Because the method does not make any attempt to control for other factors, the variability (and hence the accuracy) of the estimates is always at issue. The summary method suffers from the same sampling problems as the others, but also adds an additional wrinkle: each sample of houses is different, so it may easily be the case that a computed change in average price is primarily due to the random selection of particular houses chosen for that sample rather than to a true price change in the underlying population.

5.2 Hedonic regression methods

The idea behind the hedonic model is that the price of a house can be accurately estimated from its characteristics. The hedonic methodology was originally introduced by Court (1939); Rosen (1974) established its theoretical foundation. Other early studies were conducted by Griliches (1971), Ferri (1977), and Goodman (1978); more recent references include Thibodeau (1989), Pollakowski et al. (1991), and Meese and Wallace (1991, 1997).

The concept of aim is more problematic for hedonic methods, for the simple reason that these methods require far more assumptions. Rather than delivering a simple summary of either raw prices or raw growth rates, hedonic methods are trying to recover the functionally correct mathematical *model* of house prices. Therefore, even with idealized data, the "aim" of a hedonic model cannot be expressed in terms of the population price values without making the explicit assumption that both the functional form of the regression equation and the various parameter values are absolutely correct. Nevertheless, we argue that it is still valuable for practitioners to think about hedonic modeling in terms of fundamentals, and to consider whether the index is aiming at the desired target. For example, in many studies,

[9] However, note that because the summary is taken with prices and not growth rates, they are not in perfect alignment. For example, in the case of arithmetic mean growth rates, $\bar{\pi}_t/\bar{\pi}_{t-1} \neq \bar{\gamma}_t$.

the assumption that quality must be held constant has been considered a given. We argue that in many applications constant quality may not be so obviously desirable.

The most widely reported example of a hedonic method is probably the price index of new one-family houses sold, now jointly published by the U.S. Department of Commerce's Bureau of the Census and the U.S. Department of Housing and Urban Development. The stated goal of the price index is "to measure changes over time in the sales price of new one-family houses which are the same with respect to important physical characteristics."[10]

The advantage of the hedonic methodology is obvious. Intuitively, it seems reasonable that the price of a house should be some function of its characteristics, plus possibly some small error term to allow for remaining variation. By explicitly regressing price on a set of chosen characteristics it is possible, at least in theory, to estimate the price of a house with any characteristics at any time, and so create an index for that specific house. Whether such specificity is either necessary or desirable depends on the proposed use of the index.

The key objection to the hedonic methodology that has been raised in the literature concerns the regression step. As with any regression model, to do a good job of estimating the response variable the correct set of explanatory variables must be specified, and the correct mathematical relationships between these variables and the response variable must be guessed in advance. In general both these tasks are very difficult. Furthermore, in many cases the data have been fit using ordinary least squares, a relatively unsophisticated method. At the very least, then, the potential benefits of more specificity in the index should be weighed against the costs of imposing far more structure, an inherent requirement of hedonic modeling. Focusing on the target should help in this decision.

5.3 Repeat sales methods

Repeat sales methods focus on price *changes* rather than prices themselves, directly measuring these changes by examining only houses that have been sold at least twice. These measurements are combined in fairly intuitive ways to form estimates of the price index or growth rate in any particular time period.

As already noted, the aim of simple repeat sales methods is the geometric mean of growth rates ($\tilde{\gamma}_t$). In general this does not appear to be the implicit target of most researchers, but modifications have been suggested for delivering an estimate of the arithmetic mean ($\bar{\gamma}_t$).

One example of a repeat sales index is the Conventional Mortgage Home Price Index (CMHPI) currently produced by the Federal Home Loan Mortgage Corporation and the Federal National Mortgage Association. The CMHPI is a version of the "weighted repeat

[10]The index is scaled so that its value is 100 for 1987. A brief description is available in *Characteristics of New Housing,* a yearly additional booklet in the Current Construction Reports C25 monthly series "New One-Family Houses Sold". The index is based on a sample of about 12,000 new houses, and the actual computation requires five separate regression models, whose results are combined in a weighted average to produce the final index. The R^2s for the various regressions range from roughly 60 to 80%.

sales" methodology; the "weighted" portion of the name refers to Case and Shiller type heteroscedasticity modifications to the fundamental method.

Compared to the hedonic model, the big advantage of the repeat sales methodology is that it entirely avoids having to correctly specify the critical characteristics determining a house's value or their mathematical relationships to price. At the same time the repeat sales method is more sophisticated than the simplistic idea of taking samples and finding the central tendency of price. By using only houses that have been sold at least twice, other contributing factors to variation in price growth are controlled.

Among the objections to repeat sales, one of the most widely cited is that it is wasteful of data. In any home sales database, only some percentage of homes have been sold more than once. Obviously the extent of this problem depends on the coverage of the data used in creating the indices. For the CMHPI, for example, the number of matched pairs is roughly 25% of the number of observations in the underlying data. Note also that this objection is inherently one of efficiency: if the sampling out of sales is representative, the repeat sales method produces unbiased estimates.

5.4 Hybrid methods

Case and Quigley (1991) proposed a mixed, "hybrid" model that utilizes three different equations to apply to three different groups of transactions. Quigley (1995) extended this model to better exploit information about the underlying error structure. One equation in this framework is a hedonic regression applied to all properties that transacted only once during the sample period; one is a repeat sales regression applied to properties that transacted more than once during the sample period but had no change in property attributes (an attempt to keep quality constant); and one is a modified repeat sales regression applied to properties that transacted more than once during the sample period but had some change in property attributes.

In theory, a hybrid formulation avoids the inefficiency of using pure repeat sales since it uses information from houses that were sold only once. At the same time it uses the repeat sales idea whenever possible, thereby both exploiting the control of variation inherent in repeat sales and avoiding the problems of possible misspecification inherent in the hedonic methodology. Because of its partially hedonic-like structure, hybird models share the complicated aim of hedonic models. This, once again, raises the question of the target, and whether the benefits of the method outweigh the costs.

5.5 Comparison of methodologies by other authors

A number of articles have appeared in the real estate literature comparing the various methodologies, though none from the fundamental perspective of their aims and targets. Mark and Goldberg (1984) compared 11 models and favored five: a mean series, a median series, and three variations of the hedonic model. They rejected the repeat sales method, finding that the index values showed much smaller increases in home prices compared to some of the other models. Case et al. (1991) compared 14 models representing the repeat sales, hedonic, and hybrid methodologies. In agreement with Mark and Goldberg (1984) they found that repeat sales price estimates increased more slowly than those of the other methodologies. In disagreement with Case and Quigley (1991) they did not find any clear efficiency gains using the hybrid methodology.

Crone and Voith (1992) compared summary, hedonic and repeat sales methods. They concluded that summary methods in general were less accurate than hedonic or repeat sales methods. Among summary methods, to their surprise they found that means were better than medians. Comparing hedonic and repeat sales methods, though their results were somewhat ambiguous, they slightly favored repeat sales, in part because repeat sales was the least affected by reductions in sample size. Hosios and Pesando (1991) compared repeat sales and summary methods, finding in favor of repeat sales. Clapp and Giacotto (1992) compared their assessed value variant to pure repeat sales and preferred the assessed value method based on efficiency.

Meese and Wallace (1997) studied hedonic, repeat sales and hybrid approaches; their work is notable because they used a modern nonparametric regression technique in the hedonic estimation. They rejected repeat sales as being very sensitive to small samples, preferring either the hedonic or hybrid methodology. Finally, Gatzlaff and Ling (1994) compared median, repeat sales, hedonic and assessed value methods, using repeat sales as the benchmark. They found that all of their models produced precise estimates of the index and growth rates.

6 Conclusions

In this article, we have "gone back to basics." Instead of motivating via the more usual statistical or modeling arguments, we have instead asked more fundamental questions: what is the target? What does the repeat sales method try to deliver (with idealized data)? What does it actually deliver (with real world data)?

Using this lower level focus, we have examined the different objections to repeat sales and the various modifications of the basic method that have been proposed. Much of the disagreement over the choice of method has been shown to be due to unspoken and perhaps previously unrecognized disagreement over targets or aims. Consequently, we believe that the approach of asking simple, basic questions can be tremendously valuable to researchers and practitioners alike, and may help to ease some of the current misunderstanding and controversy over index methodology.

References

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

[2] Case, B., Pollakowski, H. and Wachter, S. (1991), "On choosing among house price index methodologies," *AREUEA Journal*, 19 (3), 286–307.

[3] Case, B. and Quigley, J. (1991), "The dynamics of real estate prices," *Review of Economics and Statistics*, 73 (3), 50–58.

[4] Case, K. and Shiller, R. (1987), "Prices of single-family homes since 1970: new indexes for four cities," *New England Economic Review*, Sep/Oct 87, 45–56.

[5] Case, K. and Shiller, R. (1989), "The efficiency of the market for single-family homes," *American Economic Review*, 79 (1), 125–137.

[6] Case, K. and Shiller, R. (1990), "Forecasting prices and excess returns in the housing market," *AREUEA Journal*, 18 (3), 253–273.

[7] Chinloy, P., Cho, M. and Megbolugbe, I. (1997), "Appraisals, transaction incentives, and smoothing," *Journal of Real Estate Finance and Economics*, 14 (1/2), 89–112.

[8] Clapp, J. and Giaccotto, 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.

[9] Clapp, J. and Giaccotto, C. (1996), "Revisions in repeat sales price indices: here today, gone tomorrow?," working paper, University of Connecticut.

[10] Court, A. (1939), "Hedonic price indexes with automotive examples," in *The Dynamics of Automobile Demand*, General Motors, New York.

[11] Crone, T. and Voith, R. (1992), "Estimating house price appreciation: a comparison of methods," *Journal of Housing Economics* 2 (4), 324–338.

[12] Ferri, M. (1977), "An application of hedonic indexing methods to monthly changes in housing prices: 1965–1975," *AREUEA Journal*, 5 (4), 455–462.

[13] Gatzlaff, D. and Haurin, D. (1994), "Sample selection and biases in local house value indices," working paper, The Ohio State University.

[14] Gatzlaff, D. and Haurin, D. (1997), "Sample selection bias and repeat-sales index estimates," *Journal of Real Estate Finance and Economics*, 14 (1/2), 3–50.

[15] Gatzlaff, D. and Ling, D. (1994), "Measuring changes in local house prices: an empirical investigation of alternative methodologies," *Journal of Urban Economics*, 35, 221–244.

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

[17] Goetzmann, W. and Spiegel, M. (1995), "Non-temporal components of real estate appreciation," *Review of Economics and Statistics*, 77 (1), 199–206.

[18] Goodman, A. (1978), "Hedonic prices, price indices and housing markets," *Journal of Urban Economics*, 5, 471–484.

[19] Goodman, A. and Thibodeau, T. (1995), "Heteroskedasticity in repeat sales house price equations," working paper, Wayne State University.

[20] Griliches, Z. (1971), "Hedonic price indexes revisited," in *Price Indexes and Quality Change* (Griliches, Ed.), Harvard, Cambridge, Massachusetts.

[21] Hill, C., Sirmans, C.F., and Knight, J. (1997), "A random walk down main street," working paper, Louisiana State University.

[22] Hosios, A. and Pesando, J. (1991), "Measuring prices in resale housing markets in Canada: evidence and implications," *Journal of Housing Economics*, 1 (4), 303–317.

[23] Mark, J. and Goldberg, M. (1984), "Alternative housing price indices: an evaluation," *AREUEA Journal*, 12 (1), 30–49.

[24] Meese, R. and Wallace, N. (1991), "Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices," *AREUEA Journal*, 19 (3), 308–332.

[25] Meese, R. and Wallace, N. (1997), "The construction of residential housing price indexes: a comparison of repeat sales, hedonic regression, and hybrid approaches," *Journal of Real Estate Finance and Economics*, 14 (1/2), 51–74.

[26] Pollakowski, H., Stegman, M., and Rohe, W. (1991) "Rates of return on housing of low- and moderate-income owners," *AREUEA Journal*, 19 (3), 417–425.

[27] Quigley, J. (1995), "A simple hybrid model for estimating real estate price indexes," *Journal of Housing Economics*, 4 (1), 1–12.

[28] Rosen, S. (1974), "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, 82, 34–55.

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

[30] Shiller, R. (1993), *Macro Markets*, Oxford University Press, Oxford, England.

[31] Thibodeau, T. (1989), "Housing price indexes from the 1974–83 SMSA annual housing surveys," *AREUEA Journal*, 17 (1), 110–117.