Quality-Adjusted Repeat-Sale House Price Indices

Author: James Bugden

La Trobe University: School of Economics

Working Paper No. 1 of 2014
Date: December, 2013

ISSN: 1837-2198
ISBN: 978-1-925085-05-1
Quality-Adjusted Repeat-Sale House Price Indices

James Bugden
La Trobe University

December 2013
Abstract

In the standard repeat-sales method, the maintained assumption is that the quality does not change between a house’s two sale dates. This assumption has been called into question. In particular quality changes from renovations carried out between a house’s two sales are not accounted for in the repeat-sales method.

In this paper, I consider four simple adjustments to the standard repeat-sales method that attempt to control for quality changes between sales: including an intercept term; including a set of neighbourhood indicators as independent variables; deflating the dependent variable appreciation rates by an estimate of renovations between sales; and dropping short-hold repeat-sales.

I find that the neighbourhood and renovations methods work best and that the standard repeat-sales house price index significantly over-estimates housing inflation by over twenty percentage points over a decade.

Keywords: Price Indices, Repeat-Sales, Renovations, Housing

JEL Codes: R31, C430
1 Introduction

If we wanted a measure of how house prices have changed over time we could simply calculate the means (arithmetic or geometric) of the house prices over the given time period. However, if we wish to use a house price index as a measure of inflation of housing we need some way to hold quality constant.

The repeat-sales method developed by Bailey, Muth and Nourse (1963) attempts to control for quality by calculating the appreciation rate of houses between its two sales. Let there be \( T + 1 \) time periods, \( 0, 1, \ldots, T \). Let \( s \) be the first sale date and \( s' \) is the second for each house \( j \). One regresses the (log of) the appreciation \( \frac{P_j s'}{P_j s} \) on to a set of \( T \) time categorical variables \( T_{jt} \) with no intercept term:

\[
\ln \left( \frac{P_j s'}{P_j s} \right) = \sum_{t=1}^{T} \alpha_t T_{jt} + \Delta \varepsilon_{js'}
\]

The coefficients on the time categorical variables \( \alpha_t \) are then exponentiated to construct a geometric house price index (HPI). Under the maintained assumption that the house does not change in quality between sales, the repeat-sale HPI measures housing inflation.

This approach has two advantages. First and foremost, it is parsimonious with respect to data requirements. All that one needs are the house prices and the dates of sales. Thus we are able to control for quality (by looking at the appreciation rates of the same house) without needing any variables on the quality of houses (eg lot size, closeness to schools). The second advantage is its simplicity. Besides requiring only two variables (the house prices and dates of sales), the regression can be estimated using simple OLS or WLS. There are three main problems associated with the repeat-sales method.

1. First, since we are only using houses that sell two (or more) times, we are discarding the majority of sales potentially making our estimates inefficient. Thus the repeat-sales method is only appropriate for large datasets. In this study of the Melbourne housing market from 1992Q1 - 2002Q2, however, we have 88,043 repeat sale houses and 31,739 units so this is not a problem.

2. Second, and related to the first point, since we are only using houses that sell two (or more) times, we may have biased estimators if houses

\( ^{1}T_{jt} = -1 \) if \( t \) is the first sale date; \( 1 \) if \( t \) is the second sale date; and \( 0 \) otherwise. See Bailey, Muth and Nourse (1963).
that sell multiple times appreciate at a different rate than houses that sell only once over the time period\textsuperscript{2}. We find that although repeat sales houses are approximately $10,000 cheaper than single-sales, they appreciate at the same rate. Therefore, the repeat sales model is apparently free of selection bias here. Moreover, this is consistent with theory: assets with identical risks should have identical returns.

3. The third problem of the repeat-sales method is that the maintained assumption that the house did not change in quality between sales is suspect (eg Hill 2011). If the quality did increase between sales then we would be overestimating housing inflation as some of the price change is due to higher quality housing.

Most of the solutions to the third problem, changed quality between sales, require data on the housing (and neighbourhood) characteristics that have changed.

One solution is to use only houses in the repeat-sales regression that have not changed in characteristics between sales. For example, Case and Shiller (1987, 1989) chose single-family houses where the number of rooms, number of bedrooms, the condition of house and modernized indicators remained unchanged (or a minor change). See also Palmquist (1980).

In the \textit{augmented} repeat-sales model, the changes in characteristics are included as independent variables (eg Bailey, Muth and Nourse 1963; Quigley 1995 and Zabel 1999).

And in the \textit{hybrid} model information on both single-sales and repeat-sales are used in both hedonic (and thus using characteristics) and repeat-sales

\textsuperscript{2}Clapp and Giacotto (1992, 1999) find that there is a sample selection bias but that it disappears when short-hold repeat-sales (observations with a short time between sales) are eliminated. They hypothesize that short-sale repeat-sale houses are more likely to have been renovated.

Likewise Steele and Goy (1997) also find a sample section bias in short-hold repeat-sales observations but attribute it to opportune buyers finding under-valued houses for re-sale. They find the first sale of a repeat sales observation sell at a 2% discount on average and the bias in the repeat-sales index to vary period to period depending on the distribution of first and second sales in the sample. They find little bias for early periods of the index but more than 3% bias for the last period. Both of the preceding papers provide a rationale for the common practice of removing short-hold sales from repeat-sale observations.

Gatzlaff and Haurin (1997) use Heckman’s selection correction model (with two lambdas one for houses that sell only once and one for houses that did not sell at all in the sample period) to correct for sample selection bias. They find the sample selection bias is statistically significant (at the 1% level) with the bias being pro-cyclical. However, they conclude that the evidence that the standard repeat-sales index magnifies the true index due to sample selection is weak.
regressions. For example, Case and Quigley (1995) run a stacked three equation regression that combines information on unchanged repeat sale houses, changed repeat sale houses and single sales all into one joint GLS estimation (See also Hill, Knight and Sirmans 1997; Englund, Quigley and Redfearn, 1999; and, for Australia, Jones 2010).

There are three problems with these approaches. First, the data on housing characteristics may not be available. For example, the dataset used in this study has no house characteristics other than location and distinguishing between houses and units. Second, if the data on housing characteristics is incomplete we will have an omitted variables problem. Third, even with all of the data we may mis-specify the functional form. The latter two problems will lead to biased and inconsistent estimates of the house price index.

We would like a method that controls for quality and yet retains the simplicity and parsimony of the repeat-sales method. In this paper we compare four simple methods that can be used to control for quality changes in the repeat-sales regression.

1. The simplest method is to include an intercept term in the repeat-sales regression. Goetzmann and Spiegel (1995) propose that much of the quality changes occur at the time of sale (atemporal appreciation) rather than over time. Thus the intercept term will pick up the changes in quality between sales.

2. A disadvantage of the intercept method is that all houses are assigned the same (atemporal) quality change. Since it is reasonable to believe that houses within a neighbourhood tend to be more similar, an alternative is to use neighbourhood indicators as independent variables in place of the intercept term. The neighbourhood indicator variables are picking up the (local) non-temporal changes in appreciation rates much like the intercept term in the Goetzmann and Spiegel model.

3. Since the quality changes of a house’s characteristics is essentially renovations, another approach is to directly include renovations. For example, Abelson and Chung (2005) control for house quality by subtracting their estimated national renovation rate (Renovations/Value of Buildings) from a median house price index. Bugden (2013) shows that the way to incorporate renovations between sale dates $s$ and $s'$ $(R_{jss'})$ into Normally an indicator variable is dropped to avoid perfect multi-collinearity. However, the repeat-sales regression has no intercept so that isn’t necessary. See Suits (1984); Halvorsen and Palmquist (1980) and Hirschberg and Lye (2001).
the repeat-sales regression is via the dependent variable:

\[ \ln \left( \frac{P_{js'}}{P_{js} + R_{jss}} \right) = \sum_{t=1}^{T} \alpha_{t} T_{jt} + \Delta \varepsilon_{jss} \]

Equivalently,

\[ \ln \left( \frac{P_{j,s'}}{P_{j,s} (1 + r_{j,ss'})} \right) = \sum_{t=1}^{T} \alpha_{t} T_{j,t} + \Delta \varepsilon_{j,ss'} \]

where \( r_{j,ss'} = \frac{R_{j,ss'}}{P_{j,s}} \) is house j’s cumulative renovation rate between sale dates s and \( s' \).

And intuitively we see that the dependent variable is the (log of the) appreciation excluding quality changes from renovations. Often data on renovations for individual houses will not be available. However, if we have an estimate of the population renovation rate for each period, we can calculate the the population renovation rate between sales \( r_{ss'} \). And it turns out that, if one uses an estimate of the population renovation rate (eg Abelson and Chung 2005), the resulting house price index is the simply the standard repeat-sales house price index deflated by (one plus) the renovation rate \( r_{ss'} \) (see Appendix).

4. Clapp and Giacotto (1992, 1999) find that there is a sample selection bias using only repeat-sales observations but it disappears when short-hold repeat-sales (observations with a short time between sales) are eliminated. They hypothesize that short-sale repeat-sale houses are more likely to have been renovated. Thus it is common to drop short-sale observations (say six months). This gives us a fourth method to control for quality changes in the repeat-sales model.

In this paper we estimate these four quality-adjusted repeat-sales house price indices. We find that the renovations and neighbourhood quality-adjusted methods make the most sense. And we find that the standard repeat-sales house price index significantly over-estimates housing inflation. Estimates of the over-estimate using the renovations and neighbourhood quality-adjusted methods are 21–24 percentage points for houses and 19–42 percentage points for units.
The primary dataset for the repeat sales model study is the PRISM database from the Department of Sustainability and Environment of the Victorian Government (and the Land Titles Office). The data ultimately comes from the Notice of Acquisition form submitted by solicitors after the sale of a property. The dataset records all property sales (residential, commercial, industrial and agricultural) in the Melbourne Statistical Division (Greater Melbourne) on a daily basis from January 1992 to September 2002 for a total of 950,317 records. Other than the location, the dataset did not have any characteristics of homes (e.g., number of bathrooms, lot size, etc.). Thus an hedonic house price index would not be feasible. However, the repeat-sales method is feasible. There were four types of variables in the dataset that were used to construct the repeat-sales house price indices:

1. The type of home
2. Location variables
3. Date of sale
4. Price of sale

2.1 Types of Home Variable

In the PRISM dataset there is a categorical variable - land use code (luc; original name luc1) that records the type of the property sold. The codes are used to distinguish various types of properties: residential, industrial and other. They range from general identifiers such as industrial land (luc = 5) and office whole building (61) to very specific ones such as abattoir (20) and even piggery (35). The land use codes range in value from one to ninety-nine, but only fifty-one occur in the dataset. Since we are interested in a house or unit price index we restricted the dataset to the land use codes (10 – 16) that are associated with houses and units. See the table below for the land use codes from the PRISM dataset that were used.

<table>
<thead>
<tr>
<th>Land Use Code</th>
<th>Type of Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Residential</td>
</tr>
<tr>
<td>11</td>
<td>Residential</td>
</tr>
<tr>
<td>12</td>
<td>Residential</td>
</tr>
<tr>
<td>13</td>
<td>Residential</td>
</tr>
<tr>
<td>14</td>
<td>Residential</td>
</tr>
<tr>
<td>15</td>
<td>Residential</td>
</tr>
<tr>
<td>16</td>
<td>Residential</td>
</tr>
</tbody>
</table>

Note: The dataset is neither time-series nor panel data: every day has multiple sales and no property has an observed price for every day. Each observation represents a dated sale of a property. A property can have several sales in the sample period: in fact this is necessary to use the repeat-sales method.
Other residential land use codes which referred to vacant land or entire blocks of units and non-residential land use codes were not included in our data. The House Block land use code (luc 1) represented 16.5% of the data in the PRISM dataset but due to uncertainty of the definition (whether each observation referred to one house or many) we did not include these observations either.

The residential land use codes not used are shown below.

Insert Table 2 here

Restricting the dataset to land use codes 10 – 16 for houses and units reduced the dataset to 769,343 observations⁵. In this paper House will refer to land use codes 10 – 13 and Unit will refer to land use codes 14 – 16. Notice almost all houses in Table 1 were classified as land use code 11: previously occupied house. It is clear that one reason for this is that house variables were frequently miscoded: e.g. given their under-representation in the dataset, new houses and terraces were likely to be frequently coded as an existing home (luc 11)⁶. That is why properties with land use codes 10 – 13 are lumped together as Houses in our study (and properties with land use codes 14 – 16 as Units). Unstably coded repeat sales – sales coded as a house in the first sale and a unit in the second or vice-versa – were deleted from the repeat-sales regression. They represented approximately 4% of the repeat-sales observations. We feel more confident that the remaining stably coded data (both sales recorded as houses or both as units) were generally coded into the correct House and Unit groups. Therefore, separate price indexes for houses and units would still be valid.

2.2 Location Variables

House addresses in the dataset are determined by the seven location variables shown in the table below:

Insert Table 3 here

Note: the street_x variable is used to differentiate 120A and 120B High Street for example. The Lotno variable, in the case of units, refers to the

⁵See Table 6 for a summary of all deletions.
⁶In correspondence with Department of Sustainability and Environment it was revealed that initially there was no code for new houses and even when it was instituted many solicitors were unaware of the change.
apartment number in an apartment building and, in the case of houses, refers to the lot number.

Since the repeat-sales method compares the same houses price over two periods we need to uniquely identify houses in the dataset. To be considered the same house, the two observations needed to match all of the address fields: street name, street number, street_x, apartment number, suburb, postcode and municipality\(^7\). An address is considered valid whether or not it has an entry for the street_x variable (an alphabetic addendum to the street number) eg both 48 King St and 50A King St are valid addresses. Moreover, houses but not units are considered to have valid addresses even if they have no lot / apt number (lotno).

2.2.1 Municipality

There are thirty-one Melbourne municipalities in the database all of which are represented in the repeat sales index. There is an average of 2,840 houses and 1,024 units in a municipality.

2.2.2 Postcode

After analysing the dataset it became apparent that in too many cases the attributed postcode did not match the suburb and municipality fields. It was decided that the suburb and municipality fields were more accurate. First, in sampling the database, the suburb and municipality fields were almost always consistent. Second, many postcode errors were obvious recording errors e.g. 3068 for 3086. And lastly, it was felt that solicitors filling out the form are more likely to remember the suburb than the postcode. To solve these conflicts we replaced the PRISM database postcode with the one listed for that suburb in the AusPost Postcode Datafile pc-full_20040928.zip\(^8\). For the suburbs not listed in the AusPost Postcode Datafile we used the postcode given in the Melways (edition 26, 1999) or failing that we accepted the postcode as given in the dataset.

Two suburbs have an extra postcode for P.O. Boxes (see table below). If the address matched for two records except for this difference they are almost certainly the same home. For this reason we substituted the Delivery Address postcode.

Insert Table 4 here

\(^7\)In our dataset I created a unique ID number for each house to facilitate matching the repeat sales. Likewise I converted the string variables Municipality and Suburb to categorical variables.

There are two Victorian suburbs with the same name in Greater Melbourne: a Hillside in municipalities Melton / Brimbank (PC 3037) and a Hillside (PC 3875) 224 km from Melbourne. Since all of these records stated Melton or Brimbank as the municipality it was assumed that all of the data referred to postcode 3037. Incidentally, Hillside PC 3037 illustrates the fact that suburbs and postcodes may belong to more than one municipality (here Melton and Brimbank). However, with the adjustments to the suburbs Dandenong South and Melbourne noted in the table above, each suburb will belong to only one postcode. There are a total of 274 postcodes in the dataset ranging from PC 3000 to PC 3984; 261 are represented in the repeat sales index.

### 2.2.3 Suburb

There are 576 suburbs in the database (467 in the repeat sales index) ranging as far as 100 km from the Melbourne Central Business District (CBD): Mount Pleasant in the Mornington Peninsula. Two kinds of corrections were made to suburb names. First, spelling mistakes were corrected: e.g. Bundoola was changed to Bundoora. Second, suburb names were standardized to conform to the AusPost names: e.g. North Fitzroy standardized to Fitzroy North. In addition, it was noticed that the suburb listed in the data as Mount actually referred to four suburbs: Mount Waverley, Mount Evelyn, Mount Eliza and Mount Martha. These were separated into their four proper suburb names based on the municipality name and postcode in the dataset. Postcode and suburb data were used only to match addresses for repeat-sale houses.

### 2.2.4 Street

There are four street variables (see the table below). However, the street type was actually recorded in the same field as the street name and the street type field (sttype) was blank. We deleted records that had no street name or street number or a street number of zero.

Insert Table 5 here

### 2.2.5 Lot number / Apartment number Variable

The lotno variable (original name lotno1) refers to the lot number or the apartment number on the plan that belongs to the property. To be conservative we treated two homes with the same street name and number as different homes if the lot number differed. For units on the same property
this is the only way to distinguish among them\textsuperscript{9}. Thus we deleted any unit addresses with missing lot numbers or a zero lot number.

Invalid street names, street numbers and lot numbers reduced the residential sales data by 6%: of the 769, 343 house and unit observations, 46, 123 had invalid addresses (e.g. an address with no street number) and were discarded leaving 723, 220 observations.

To summarize, a house (or unit) was considered the same house (unit) in the dataset if it matched all of the municipality, postcode, suburb, street name, street number, street\_x and lot number fields. We deleted observations with no street name or street number (or a street number of 0). We also deleted any units with no lot number (or a lot number of 0).

\subsection*{2.3 Date Variable}

The date variable (original name condate) is of daily frequency and represents the contract date\textsuperscript{10}. The contract date is considered a better measure than the settlement date as it is closer to the time the market price is determined\textsuperscript{11}. Although the dataset covers the time period from January 1, 1992 to August 27, 2002, notices of acquisition in Victoria can be submitted up to one-hundred and twenty days after the date of settlement and so the last four months are incomplete.

But although it is misleading for sales, the number of residential sales in June and July, 2002 is large enough (733 and 204 respectively) to get a sufficient number of repeat sales to estimate prices. For August and September there are too few residential sales (32 and 3 respectively). For this reason we restricted our sample to 1992Q1 to 2002Q2.

Therefore, our study comprises a total of 42 quarters (126 months). We also deleted any observations with both sales in the same quarter as two sales in the same time period cannot be properly coded using the repeat-sales method\textsuperscript{12} (and they could be double entries of the same sale as well).

\textsuperscript{9}It is assumed that any sub-dividing of properties will lead to an addition of lot numbers to distinguish the now two (or more) properties. Thus treating houses with different lot numbers as different houses should avoid the changes in quality due to sub-division.

\textsuperscript{10}Confirmed via email correspondence with Department of Sustainability and Environment.

\textsuperscript{11}ABS 6417.0 Renovating the Established House Price Index, November 2005 p.4.

\textsuperscript{12}In the repeat sales method there is a time indicator variable for each period that is coded $-1$ if first sale, 1 if second sale and 0 if no sale that period. Thus two sales in the same period would be coded $-1 + 1 = 0$, the same as a house that does not sell that period.
2.4 13.4. Price Variable

The (nominal) price variable (original name consider) is the total amount the buyer paid for the home. Thus it would include stamp duties and (for new homes) the GST\textsuperscript{13}. We ignored uneconomic transactions, defined to be a sale price less than $10,000. This reduced the sample by only 265 records. Ignoring sales with a price less than $1,000 would have reduced the sample by 31. Restricting our data to economic transactions of houses with valid addresses from 1992Q1 - 2002Q2, left 713,975 home sales. Of these, 523,106 were house sales and 190,690 were unit sales. Of the 523,106 house sales, 169,401 were houses that sold at least twice in the sample period. Two house sales make one repeat sale observation so, after eliminating repeat sales within the same quarter, we were left with 88,043 repeat sale house observations. Likewise, the 190,690 valid unit homes resulted in 31,739 repeat sale unit observations.

2.5 Summary of Deletions

A summary of the deletions is given in the table below.

Insert Table 6 here

If we are constructing a monthly index there would be slightly more repeat sales observations pairs as there must be more dropped dual sales in the same quarter than dual sales in the same month. For example a repeat sale observation with the first sale in January and second in March would be in the same quarter and thus dropped in a quarterly index, but in a different month so not dropped in a monthly index.

2.6 Renovations Variable

Besides postcodes (from the Australian Post), the only other variable that we use that is not in the PRISM dataset is the renovation variable. We use Abelson and Chungs (2005) estimates of the overall Australian rate of renovation (ie the value of renovations as a proportion of the value of the housing stock) over the sample period.

To construct the renovation rate, they first construct a series for the value of the housing stock by first taking the value from the 1991 Population and Housing Census and then estimating the other years by assuming a 1.75\% increase.

\textsuperscript{13}Correspondence via email with Department of Sustainability and Environment states "The Consideration is the total cost the person paid for the property".

10
change in the quantity of the housing stock per year (and estimating the nominal dwelling prices). Once they have the housing stock series, they calculate the Expenditure on Alterations and Additions (ABS Cat 5204.0) as a proportion of the Value of the Housing Stock to generate the renovation rate. To convert these (one plus) annual renovation rates $1 + r$ to a quarterly measure, we used geometric averaging: (one plus) quarterly renovation rate $1 + r_q = (1 + r)^{\frac{1}{4}}$. The estimated renovation rate by Abelson and Chung (2005) varied by year but was approximately 1% per year, and yielded a cumulative rate of approximately 11% over our sample period 1992Q1 to 2002Q2. Although the yearly renovation rate is fairly constant, recall dependent variable in the renovation-correction method (method 3) is deflated by the amount of the estimated renovation between sales. So homes with longer (shorter) times between sales will have more (less) cumulative renovations discounted from the homes appreciation rate.

2.7 Choice of Time Period: Monthly or Quarterly

The repeat sales technique estimates the log of the cumulative appreciation rate and the OLS estimator is a (complicated) weighted estimator of the sample means $\bar{\alpha}$ (of the log of the cumulative appreciation rate) from every pair of sale dates: a sale date pair is the first sale date and the second sale date. For example, if $T = 2$ then the repeat sales estimators of the log of the cumulative appreciation rate is

$$\hat{\alpha} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} w_{11}^* \bar{\alpha}_{01} + w_{12}^* (\bar{\alpha}_{02} - \bar{\alpha}_{12}) \\ w_{21}^* \bar{\alpha}_{02} + w_{22}^* (\bar{\alpha}_{01} + \bar{\alpha}_{12}) \end{bmatrix}$$

where

$$w_{11}^* = \frac{n_{01}(n_{02} + n_{12})}{n_{01}(n_{02} + n_{12}) + n_{02}n_{12}}$$

$$w_{12}^* = \frac{n_{02}n_{12}}{n_{01}(n_{02} + n_{12}) + n_{02}n_{12}}$$

$$w_{21}^* = \frac{n_{02}(n_{01} + n_{12})}{n_{01}(n_{02} + n_{12}) + n_{02}n_{12}}$$

$$w_{22}^* = \frac{n_{01}n_{12}}{n_{01}(n_{02} + n_{12}) + n_{02}n_{12}}$$

The subscripts on the number of sales $n$ and the sample mean $\bar{\alpha}$ refer to the period of the first and second sale respectively. For example $\bar{\alpha}_{02}$ is the sample mean of the logged cumulative appreciation rate of all houses that sell first in the base period ($t = 0$) and then again in period ($t = 2$). We see that the weights depend on the number of observations for all sale date pairs. For example, suppose there are no repeat sale observations with the first sale in the base period 0 and the second sale in period 2 ie $n_{02} = 0$. The weights for

\footnote{See Bailey Muth and Nourse (1963).}
the repeat sales estimator $\hat{\alpha}_1$ reduces to $w_{11}^* = 1$ and $w_{12}^* = 0$. Now suppose instead $n_{01} = 0$. Now $w_{11}^* = 0$ and $w_{12}^* = 1$. Consequently, volatility in the number of observations for each sale date pair will lead to unstable estimates of the appreciation rates.

This is more likely to be a problem the finer the time period (i.e., more time periods in a given span of time) as each time period will have on average fewer observations. In our dataset we have $T + 1 = 42$ quarters which yields $\frac{42 \cdot 41}{2} = 861$ possible sale date pairs e.g. (Q17, Q31). With 88,043 repeat sale observations on houses and 31,739 on units each sale date pair would have, on average, 102 observations for houses and 37 for units.

If we used monthly data ($T + 1 = 126$ months in our study) we would have $\frac{126 \cdot 125}{2} = 7,875$ possible sale date pairs: almost a magnitude more than the quarterly sale date pairs. Each sale date pair would have, on average, 11 observations for houses and only 4 for units. Thus random changes in sales would have a bigger effect on estimates of unit appreciation rates using monthly periodicity.

For example, suppose that for purely random reasons a house may sell in January or February. Which month it sells on would affect our estimates if we used a monthly time period as the number of observations in the sale date pairs involving that January (and also February) would change thus changing the weights. And note that the estimated appreciation rates for every period would change as each is a function of all of the sale date pair observations. If we used a quarterly time period, however, our estimates would not be affected.

Volatility is not the only problem. Sommervoll (2006) constructs a Monte Carlo simulation of 100 samples each with 1,000 repeat sales. He finds that samples with finer-grained time periods are not only more volatile (i.e., higher standard errors around the true price index) but are also more likely to seriously over or under-estimate the true price index. This latter effect is due to the fewer number of observations per time period which makes each observation more influential on the estimate. On the other hand Sommervoll points out that a coarser-grained time period is less likely to track turning points in the price index (local maxima and minima). Moreover, as the base period price index is normalized to $e^0 = 1$, any growth in price index within the base period will not be picked up in estimation. Thus coarser-grained time period estimation is more likely to under-estimate appreciation rates.

We were not interested in identifying the turning points in the price index as much as estimating the price index per se so we decided to use a denser data set with the coarser-grained approach (quarterly instead of monthly). However, we did check the estimates of the price index using both coarse-grained (quarterly) and fine-grained (monthly) time periods. Given the rel-
ative sparseness of the data on units this suggests that monthly / quarterly choice may be more of an issue for units than houses. And that is indeed what we find. In checking the estimated monthly price indices, we found that it produced nearly identical estimates to the quarterly series for houses, while for units the monthly Unit Price Index was typically 5% higher than the equivalent quarterly one. This may be the base period normalization problem mentioned above. However, the monthly index dropped fewer observations than the quarterly one (see footnote 12). Thus the difference in estimates for units may also be due to the sensitivity of the weights to changes in the number of observations per sale date pair.

3 Summary Statistics

The graphs of prices and sales in this sections refer to all valid homes; that is, houses and units with a valid address and that sell for at least $10,000. There are 523,285 such houses and 190,690 such units.

3.1 Prices

Over the entire sample period (1992Q1 - 2002Q2) house prices ranged from the minimum we allowed, $10,000, to $10,000,000 and unit prices ranged from $10,000 to $3,950,000. The overall distribution of house and unit prices looks lognormal (see Figures 1 and 2).

The (one plus) appreciation rate of home $j$'s price from the time of the first sale $s$ to the time of the second sale $s'$ is $P_{js} / P_{js'}$. The log of this appreciation rate is the dependent variable in the repeat-sales regression. Figure 3 shows the arithmetic average prices of valid houses and units by quarter over 1992Q1 - 2002Q2.

There are two things to note. First, prices of both houses and units were relatively constant from 1992 to 1997 but then rose significantly from 1997 to the end of the sample, 2002Q2. Second, as expected, house prices are higher than unit prices around $20,000 for much of the sample. This difference probably reflects a difference in quality broadly defined; in particular the size of the home.
3.2 Sales

Just as house and unit prices increased markedly after 1997 so did sales (see Figure 4).

Relative to the mean, unit sales over time were more variable than house sales: the coefficient of variation is 0.23 for units and 0.17 for houses\textsuperscript{15}.

3.3 Repeat Sales

The repeat sales method only uses homes that sell two or more times. As mentioned in the data section, there are 88,043 house and 31,739 unit repeat sale observations (pairs of sales). Of these, approximately three-quarters only sold twice and over 99% sold from two to four times whether houses or units.

Repeat sale houses are less expensive compared to houses that sell only once over the decade – approximately $10,000 in this sample (see Figure 5). The same result holds for units. This difference is usually attributed to a difference in quality. For example, starter homes are less expensive homes that people buy initially, but then sell when they can afford the actual home they want. Thus these homes tend to sell more often and hence these less expensive homes may be over-represented in the repeat sale homes.

As a consequence, it is then argued, repeat sale homes suffer from a sample selection bias which affects the estimated house price index. However, this misses an important point. Unlike the hedonic regression which estimates average house prices, the repeat sales method estimates average appreciation rates of house prices. And crucially, we find that, despite being cheaper, repeat sale houses appreciate at rates similar to single selling houses in this dataset (Figure 6 shows the similar per-period appreciation rates). The same holds true for units. And homes that sell three or more times also appreciated at similar rates as single or double-sales. Thus that sample selection bias does not appear to be a problem here.

\textsuperscript{15}CV = \frac{\text{standard deviation}}{\text{mean}}. The last quarter was dropped as not all sales for that quarter were recorded at the time of this dataset's construction.
4 Results

4.1 A Note on the Error Terms

The OLS model assumes that the error terms are independently and identically distributed. We considered four variations of the error term:

1. Case and Shiller’s Weighted Repeat-Sales (WRS) model
2. Quigley and Van Order’s Weighted Repeat-Sales (WRS) model
3. Robust standard errors
4. Inter-dependent error terms of observations of the same house

4.1.1 Weighted Repeat-Sales (WRS) models

Case and Shiller hypothesised that heteroskedasticity is a problem for repeat sales regressions. Houses with a long time between sales are more likely to have changed quality and are thus less precise estimates of inflation. This means the variance of the error term should increase with the time between sales.

In Case and Shiller’s three-stage Weighted repeat sales (WRS) method the square of the residuals from the standard repeat sales model (1st stage) were regressed on the number of days between sales (2nd stage)\(^{16}\). The inverse of the fitted residuals from this regression were used as weights in the new repeat sales regression (3rd stage).

The results of the second stage are shown below for houses.

Insert Table 7 here

The variance, as proxied by the square of the residuals, does appear to depend on the time between sales (at least at the 5\% level). However, the negative coefficient implies that houses that sell a long time apart are more precise estimates of the appreciation rate; this is the opposite of Case and Shiller’s theory and motivation for the WRS method.

\(^{16}\)Graddy et al (2012) develop an estimator similar to Case and Shiller’s three-stage weighted repeat sales estimator but with weights constructed in the second stage by regressing the squared residuals from the first stage regression on indicator variables representing the length of the holding period for each asset.
We then considered Quigley & Van Order’s version of WRS method which added a squared term of the time between sales to the 2nd stage regression to capture non-linearity. The results of the second stage are shown below.

Insert Table 8 here

The squared time between sales term is significant (at virtually any level). We can see the non-linear relationship by looking at the fitted values of the OLS squared residuals from the 2nd stage (see Figure 7).

Insert Figure 7 here

The relationship between the variance of the error term and the time between sales is quadratic which is why Case and Shillers result did not make sense. We see now that houses that sell a long time apart are less precise estimates of the appreciation rate just as Case and Shiller hypothesised. But we also see that houses that sell a short time apart are also less precise estimates of the appreciation rate. Units (apartments) showed the same pattern. For this reason we preferred Quigley & Van Order’s WRS to Case and Shiller. However, the OLS and WRS standard errors were virtually identical. The efficiency gains of the WRS over OLS are limited – a reflection of the large sample size.

4.1.2 Robust standard errors

We then considered, instead of weighted repeat sales, robust standard errors (also known as Huber and White standard errors). These are able to correct for other forms of heteroskedasticity not captured by the time between sales. Not only are they valid for arbitrary heteroskedasticity, they also are valid whether or not there is heteroskedasticity (Wooldridge, 2002 p. 57). Moreover, efficiency losses from not specifying the form of the heteroskedasticity is less of a problem given the very large number of observations we have.

We found that the robust standard errors of the coefficients were smaller than the OLS and WRS standard errors. However, whichever method we used, (OLS, WRS or robust OLS), there was virtually no difference in the estimated house price index and a modest difference in the unit price index with WRS estimates slightly lower than OLS or robust OLS.

4.1.3 Inter-dependent Error Terms

Another possible concern is that the error terms may not be independent of each other. In particular, for a house that sells three times, the appreciation
rate between the first and second sale may be related to the appreciation rate between the second and third sale.

However, the dependence is not obvious a priori. Is it the case that there are houses that are particularly fashionable / unfashionable over a sustained period of time for whatever reason? If so, their appreciation rates would be under-estimated / over-estimated and their error terms will be positively correlated. Or is it the case that a higher (lower) appreciation from the first to second sale indicates the second buyer overpaid (underpaid) and thus the second appreciation rate is likely to be lower (higher)? In this case there is a negative relationship between the error terms. This is an interesting question for future work.

The econometric / data package Stata has an option with the robust estimator, (called \texttt{cluster()}), to correct for any specified dependency of the error terms. Essentially it groups together those observations one believes are dependent and then assumes independence across groups just like OLS does across observations\textsuperscript{17}. Given the above discussion, we chose to group together repeat sale observations of the same house. Recognizing the dependency of appreciation rates of the same house led to a modest increase in the standard errors. We considered these the appropriate standard errors.

Finally, the estimated coefficients for houses had lower standard errors than for units. This is likely due to the difference in sample sizes (88,043 for houses; 31,739 for units). Other than the lower cumulative appreciation rate and higher standard errors units behave much like houses in our sample.

4.2 Standard Repeat-Sales

The OLS estimation for both houses and units of the standard version of the repeat sales house price indices is shown in Figure 8.

Insert Figure 8 here

Over the sample 1992Q1 to 2002Q2 (42 quarters), the standard repeat-sales model estimates that houses appreciated 115\% (HPI\textsubscript{42} = 2.150343) and units 94\% (UPI\textsubscript{42} = 1.94166). Not only are units (apartments) cheaper, they also appreciated less than houses over the decade. By comparison, general inflation as measured by the ABS Melbourne CPI increased only 23\% over the same period\textsuperscript{18}.

\textsuperscript{17}See Stata 9 Reference Manual R Z, Regress.
\textsuperscript{18}Melbourne CPI from ABS 6401.0 Consumer Price Index, Australia, Table 1B; CPI Melbourne (Quarterly) (1989 90 Base Year). As the GST was introduced in 2001Q3.
4.3 Quality-Adjusted Repeat-Sales House Price Indices

A summary of the quality-adjusted regressions is shown in Figure 9 and the table below.

Insert Figure 9 here

Insert Table 9 here

4.3.1 Renovations-Adjusted Repeat-Sales model

To correct the standard repeat-sales model for changes in quality, we first deflate the dependent variable \( \frac{P_{js}}{P_{js}} \) by the estimated population renovation rate between sales \( 1 + r_{ss} \). As mentioned in the Introduction (and shown in the Appendix) this is equivalent to deflating the standard repeat-sales house price index at time \( t \) (HPI\(_t\)) by \( 1 + r_t \) (the cumulative renovation rate to time \( t \)). Since (one plus) the cumulative renovation rate over the sample (42 quarters) is \( 1 + r_{Q42} = 1.106674 \) we have a renovations-adjusted repeat-sales house price index over the entire sample of \( r\text{-HPI}_{42} = \frac{HPI_{42}}{1+r_{Q42}} = \frac{2.150343}{1.106674} = 1.943069 \) and unit price index of \( r\text{-UPI}_{42} = \frac{HPI_{42}}{1+r_{Q42}} = \frac{1.94166}{1.106674} = 1.754501 \).

Notice that a cumulative renovation rate of 11% reduced the cumulative price index by 21 percentage points for houses (2.15 to 1.94) and 19 percentage points for units (1.94 to 1.75). Thus the renovation correction to the standard repeat-sales price index is significant even though the estimated renovation rate is only approximately 1% per year.

4.3.2 Intercept Repeat-Sales model

An alternative to directly including a renovation rate in the repeat-sales regression is to include an intercept term instead (Goetzmann and Spiegel, 1995):

\[
\Delta p_{j,s,s'} = \alpha + \sum_{t=1}^{T} \alpha_t T_{j,t} + \Delta e_{j,s,s'}
\]

The intercept represents the non-temporal appreciation rate in house prices. They argue that home-owners renovate around the time of sale and

ABS recommends adjusting their CPI downwards by 3% of the CPI from that point (ABS Special Article: Measuring the impact of the new tax system on the September Quarter 2000 Consumer Price Index, p.2). The unadjusted increase in the Melbourne CPI over the sample period was 26%.
thus renovations are captured by the intercept term. Using this methodology we found house prices (adjusted for quality) increased about 80% over the sample (1992Q1 - 2002Q2) compared to the 94% increase we found using renovation model. Likewise, the intercept model estimates units increased 48% over the sample compared to the renovation model estimate of 75%.

Why are they significantly lower than the estimates in the renovations-adjusted model? Shiller (1993) finds that in the presence of heteroskedasticity the constant term biases the price index downwards (consistent with our findings).

Note if we exponentiate the estimated intercept term, we get a cumulative non-temporal appreciation of approximately 9% for houses and 13% for units. If we interpret this non-temporal appreciation as due to renovations at the time of sale (as Goetzmann and Spiegel do) the renovation rates of 9% and 13% seems comparable to Abelson and Chungs estimate of 11%. However, in the renovation model homes are deflated by the cumulative renovation rate \textit{between sales}. Only homes that have the first sale in the base period \((t = 0)\) and the last quarter \((t = 42)\) will be have their appreciation rate deflated by the entire 11%. Most homes will be deflated by much less.

For example, the average repeat-sale house in the sample has roughly 14 quarters between sales. If, for simplicity, we say the renovation rate is a constant 1% per year, this translates into a quarterly renovation rate of \((1.01)^{\frac{2}{14}} = 1.0025\) and a cumulative renovation rate over 14 quarters of \((1.01)^{14} = 1.035\). Thus the average house’s appreciation will only deflated by about 3.5% compared to 9% for every house in the intercept model. Interestingly, in McMillen and Thorsnes (2006) study of the Chicago housing markety, they attempt to correct for renovations by excluding houses that have building permits in their sample; and they find that the standard repeat-sales house price index is about 4.4% higher over a decade than estimates where houses with building permits are removed. This is similar to our estimate above of 3.5% over a similar time span.

The renovation rate model has two advantages over the intercept model. First, if data on individual house renovations are available, the renovation rate model can accommodate this better information. Second, even in the case where we using population renovation rates (as we do here), the renovation rate model allows the renovation rate to vary as the time between sales varies whereas the in the intercept model the implied renovation rate is the same for all houses no matter how long (or short) the time between sales.
4.3.3 Neighbourhood Repeat-Sales model

We then controlled for neighbourhood quality by including municipality indicator variables into the repeat-sales regression. Normally one drops one indicator variable to avoid perfect multi-collinearity. However, the repeat-sales regression has no constant so we don't need to drop a neighbourhood indicator. The neighbourhood indicator variables are picking up the non-temporal appreciation much like the intercept term in the Goetzmann and Spiegel intercept model. However, unlike the intercept model this non-temporal appreciation can now vary by municipality. The time indicator variables remain the overall (logged) appreciation rates for Melbourne (but controlled for house and neighbourhood quality).

A Wald test on the municipality indicator variables indicates that they are jointly significant, hardly surprising given our sample size. (houses: $F(31, 76837) = 413.42$ prob $> F = 0$). But they are practically significant as well: the highest municipality coefficient for the house regression, Port Phillip, is $0.2187023$: 34% of the cumulative Melbourne logged house appreciation rate $0.6449832$. Exponentiating, we get a Melbourne cumulative appreciation rate of $e^{0.6449832} = 1.91$ and Port Phillip $e^{0.6449832 + 0.2187023} = 2.37$: 46 percentage points higher. We also plotted the fitted values of the regression (the log of the appreciation rate) against the residuals. The residuals do not exhibit any trending pattern or non-linearities, suggesting a well-fitted model.

Controlling for quality with municipality indicator variables, house prices rose 91% over the decade (1992Q1 - 2002Q2) compared to 94% with the renovations model. Likewise, unit prices rose 52% over the decade compared to 75% in the renovations model. The two models give comparable estimates for houses but the neighbourhood model estimate of the unit price index is much lower. It may be that units experience a greater rate of renovation than houses and thus using the same renovation rates for houses and units the renovation model over-estimates the unit inflation. However, municipalities cover a large area that may encompass many different neighbourhoods so this may be too crude of a measure to capture neighbourhood effects.

In comparing the various models (standard, renovations, neighbourhood and intercept) we cannot use $R^2$ as the standard and renovations models

---

19Note: we could not do this for the standard or renovations repeat-sales models as they have no intercept term. Here, since we are not dropping any neighbourhood indicator variables, the neighbourhood indicator variables collectively act as the intercept.

20It follows that the renovation model under-estimates house inflation. However, as houses make up the large majority of sales (74% of repeat-sales in our dataset) there is a smaller effect on the house estimate: 91% versus 94%.
have no intercept. However, using the root of the mean squared error the neighbourhood model provides the best fit.

4.3.4 Removing Short Sales

The fourth simple method to control for quality changes in the repeat-sales model is to drop observations with a short time between sales. This is based on the idea that homes that are flipped (bought and sold in a short period of time) are more likely to have been renovated. Dropping observations with a time between sales of one or two quarters only reduced our cumulative house price index by 1% (houses or units) compared to the standard repeat-sales model. Since this is much less of an effect than the renovations or neighbourhood repeat-sales models, this suggests significant quality changes are not limited to short sales. However, the root mean squared error for the unit price index - but not houses - was better than any of the other models. This suggests that dropping short-hold repeat sales is useful for unit price indices.

4.3.5 A Comparison to the ABS House Price Index

Finally we compare our quality-adjusted repeat-sales House Price Indices to the Australian Bureau of Statistics (ABS) index. The ABS index is not a repeat sales index, but is constructed from a stratified sample that controls for quality by neighbourhoods and house type. Moreover, the ABS House Price Index is an estimate of the median and the repeat sales method is an estimate of the geometric mean. However, for lognormal distributions, the expected value of the median and the geometric mean are the same\textsuperscript{21} so no adjustment is necessary.

The ABS estimated house price index increased 122\% from 1992Q1 to 2002Q2. In comparison, our standard repeat sales house price index increased by 115\%. This 8\% difference over a decade is relatively modest. Hendershott and Thibodeau (1990) find that the U.S. National Association of Realtors median house price index is approximately 2\% higher per annum than a repeat-sales HPI. The smaller difference we see is probably a reflection of the ABS attempt to control for quality through stratification.

However, if we attempt to control for quality changes, the difference is more significant. The renovations-adjusted repeat-sales house price index increased 94\% and the neighbourhood-adjusted house price index increased 91\%. Thus we attribute a significant amount of the ABS 123\% increase in house prices to quality changes.

5 Conclusion

The standard repeat sales model does not do a good job estimating house price inflation as it includes appreciation from quality changes between the two sale dates. We presented four simple ways to adjust the repeat sales model to control for quality changes. Two of them, deflating the standard repeat sales HPI by estimated renovation rates or incorporating neighbourhood indicators as independent variables, worked the best. Controlling for changes in house quality using renovations or the neighbourhood indicator variables we find that over 1992Q1 to 2002Q2, Melbourne house prices increased 91 – 94% in contrast to 115% estimated by the standard model and 122% with the ABS index. Controlling for quality has a significant effect on house price indices. We would recommend deflating the ABS house price index by approximately 1% per year for renovations.

We found little evidence of sample selection bias in repeated homes. Houses that sell two or more times appreciate at a similar rate to single sale homes (both unadjusted for quality). Indeed homes that sell three or more times also appreciated at similar rates.

6 Appendix

The following is drawn from Bugden (2013).

Recall that with the renovations quality-adjusted repeat-sales model we deflate the appreciation \( \left( \frac{P_{j,s'}}{P_{j,s}} \right) \) by the amount of renovations between sales:

\[
\ln \left( \frac{P_{j,s'}}{P_{j,s} (1 + r_{j,ss'})} \right) = \sum_{t=1}^{T} \alpha_t T_{jt} + \Delta \varepsilon_{j,ss'}
\]

If we do not have data on individual house renovation rates between sales \( (r_{j,ss'}) \) but use a population renovation rate between sales \( (r_{ss'}) \) we have

\[
\ln \left( \frac{P_{js'}}{P_{js} (1 + r_{ss'})} \right) = \sum_{t=1}^{T} \alpha_t^* T_{jt} + \Delta \varepsilon_{jss'}
\]

Since \( T_{jt} \) is \(-1\) if \( t \) is the first sale (ie \( t = s \)), \( 1 \) if \( t \) is the second sale (ie \( t = s' \)) and zero otherwise this equation is equivalent to

\[
\ln \left( \frac{P_{js'}}{P_{js} (1 + r_{ss'})} \right) = \alpha^*_s - \alpha^*_s + \Delta \varepsilon_{jss'}
\]
The resulting renovations-adjusted repeat-sales House Price Index $HPI^*_t = e^{\alpha_t^*}$ is simply the standard repeat-sales House Price Index deflated by the (one plus) population cumulative renovation rate. To see this, first note that since we are using the population renovation rate, the renovation rate will be independent of the observation $j$ so the above renovations-adjusted repeat-sales equation is

$$\ln \left( \frac{P_{j,s'}}{P_{j,s}} \right) = \alpha^*_s - \alpha^*_s + \ln \left(1 + r_{ss} \right) + \Delta \epsilon_{j,ss'}$$  \hspace{1cm} (1)

Recall the standard repeat-sales model is\footnote{As the only difference between the standard and (population) renovations-adjusted model is the constant term $\ln \left(1 + r_{ss} \right)$, the error terms are the same.}

$$\ln \left( \frac{P_{j,s'}}{P_{j,s}} \right) = \sum_{t=1}^{T} \alpha_t T_{jt} + \Delta \epsilon_{j,ss'} = \alpha^*_s - \alpha_s + \Delta \epsilon_{j,ss'}$$  \hspace{1cm} (2)

Taking the difference in equation 1 and equation 2 we have $\alpha^*_s - \alpha^*_s + \ln \left(1 + r_{ss} \right) = \alpha_s - \alpha_s$. Noting that there is no $T_{jt}$ for $t = 0$ (so $\alpha^*_0 = \alpha_0 = 0$) the cumulative appreciation from time $s = 0$ to time $s' = t$ is

$$\alpha^*_t + \ln \left(1 + r_{0t} \right) = \alpha_t \ \text{ie} \ \alpha^*_t = \alpha_t - \ln \left(1 + r_{0t} \right)$$

The renovations-adjusted repeat-sales time coefficients are the same as the standard repeat-sales time coefficients less the log of the cumulative renovation rate to time $t$. It follows that for the house price index we have

$$HPI^* = e^{\alpha^*_t} = e^{(\alpha_t - \ln \left(1 + r_{0t} \right))} = \frac{e^{\alpha_t}}{1 + r_{0t}} = \frac{HPI}{1 + r_{0t}} \ \text{ie} \ \frac{HPI}{HPI^*} = 1 + r_{0t}$$

The renovations-adjusted repeat-sales house price index is equal to the standard one deflated by (one plus) the population renovation rate. Therefore, a simple way to have a quality-adjusted repeat-sales price index is to deflate the standard repeat-sales price index by an estimate of the population cumulative renovation rate.

References


