Private Residential Price Indices in Singapore: A Matching Approach

Yongheng Deng
Daniel P. McMillen
Tien Foo Sing

June 2011
We use a matching procedure to construct samples of private residential sales in Singapore for January 1995 to May 2010. Though the matching approach is similar to a repeat sales estimator in pairing each sale with the sale of a comparable property, sample sizes are much larger because the matched properties are not constrained to be identical in each period. An advantage of the matching procedure is that it makes it easy to characterize changes in the full distribution of quality-adjusted sales prices, rather than just the means. We find that the distribution of sales prices shifted much further to the right at high prices than at lower prices for 1995-2010, and this pattern is particularly evident in the boom periods of 1996 and 2005-2007. The variance of the sales price distribution increased significantly during boom periods.

* The authors thank Maggie Hu for outstanding research assistance and Liu Bo for assistance in the data collection.
1. Introduction

The standard repeat sales approach for constructing price indices attempts to control for quality changes by restricting the sample to homes that have sold at least twice over the sample period. If the explanatory variables and the coefficients of the underlying hedonic price function are constant over time, then the repeat sales approach provides unbiased estimates of pure time effects without having to include any controls for structural or neighborhood characteristics in the estimating equation.¹ This potential advantage of the repeat sales estimator over hedonic estimation comes at the cost of a large reduction in the sample size. A lack of data often prevents repeat sales price indices from being estimated for small geographic areas or for emerging markets with limited repeat sales. In addition, it has been documented that repeat sales price indices may suffer from revision bias.² The potential non-randomness of sample houses sold more than once creates another problem for the repeat sale price indices.³

As mean-based estimators, neither the repeat sales nor the hedonic approach is well suited to analyzing changes in the full distribution of sales prices over time. Both approaches are based on an implicit assumption that the distribution of the natural logarithm of sales price simply undergoes a parallel shift across time periods, i.e., prices change by the same percentage at both the upper and lower ends of the sales price distribution. In contrast, McMillen (2008) finds that the upper end of the sales price distribution shifted farther to the right than the lower end of the distribution between 1995 and 2005 in Chicago.

¹ The maintained hypotheses underpinning the repeat sales estimators are discussed in Case and Quigley (1991), Meese and Wallace (1991, 1997), and Englund, Quigley and Redfearn (1998 and 1999).
² See Deng and Quigley (2008) for a discussion on the magnitude and bias of the revision in estimating repeat sales price indices.
³ Englund, Quigley and Redfearn (1999) found that “starter” houses were sold more frequently than more expensive houses in Sweden. Gatzlaff and Haurin (1997), however, showed that changing economic conditions cause variations in offer and reservation prices, which in turn affect the composition of houses sold in a non-random manner.
McMillen (2010) proposes a matching procedure to construct sales price indices for any target percentile of the sales price distribution. The construction of a house price index has a direct parallel to estimating average treatment effects: what is the expected sales price of an otherwise identical property if it sells at time \( t+s \) rather than at time \( t \)? In the matching literature, matched samples are constructed by pairing each treated observation with a comparable control observation (or vice versa). In our paper, matching is based on the propensity scores of houses with observed housing and location characteristics being sold between the control (base) and treatment periods. The repeat sales model is an extreme version of a matching estimator in which each observation is paired with its closest neighbor – itself. Averages over many observations do not require such an extreme form of matching to lead to unbiased estimates of treatment effects. Outliers and unrepresentative sales are removed from the sample\(^4\) to further reduce biases in the matching process. Furthermore, the matching procedure can effectively mitigate the index revision bias because the indices estimated using matching procedure are based on pairing samples constructed through comparable control observations rather than specific individual transactions.

Wang and Zorn (1997) show that the repeat sales estimator is equivalent to a simple period-by-period comparison of means if the number of sales is the same in each time period. Since matched samples can be constructed by pairing each base period observation with a comparable property from each subsequent period, the repeat sales estimator can be simplified to period-by-period differences between the mean in the base period and at all other times. There is no reason why this comparison must be limited to the mean value. It is straightforward to construct differences in medians, or to focus on any other percentile of the price distribution. The matched sample approach makes it easy to construct price indices for all percentiles of the sales price distribution. Also, matched samples help to reduce problems associated with non-random samples,

\(^4\) Older developments sold collectively for the redevelopment potential are excluded in our study. The exclusion of these small fractions of samples is not likely to cause selection bias problems.
and thus help to ensure that the implied changes in price distributions the true underlying market conditions. In this paper, we use a matched sample approach to construct price indices for sales of private residences in Singapore for 1995-2010. Singapore had two episodes of extremely rapid price appreciation during this period – a relatively short boom in 1996 and an extended run-up of prices over 2005-2007. All three approaches to price index construction – standard repeat sales, hedonic estimation, and the matched sample approach – indicate that the average price of a quality-constant home increased significantly during this time, with the hedonic and matched sample results being closer to one another than to the repeat sales model and its more restrictive sample of homes. Quantile estimates for the matched samples suggest that the price distribution shifted much farther to the right at high sales prices during these boom periods, a result that is consistent with McMillen’s (2008) findings for Chicago. These results imply that the variance of sales prices increased markedly during times of rapidly appreciating sales prices.

2. Singapore’s Housing Market

Singapore’s housing market has a dual-market structure that combines central planning and a private market. Since gaining independence in 1965, the government of Singapore has used its comprehensive housing program to directly provide affordable public housing for close to 90% of the population. In addition, the government sells land to private developers who then build high-rise and high-end condominium housing to sell at market rates to wealthy households. Introduced in the 1980s, high-rise condos with shared facilities introduced have become popular among middle-income young professional couples. Private developers sell build condos and sell strata title units to individual owners. Strata-titled owners share joint interest on land and generally have 99-year leasehold tenures. The private housing stock increased by more than 3.5 times over the last 20 years.

---

5 The state retains reversionary interest on 99-year leasehold lands. Though there are still pockets of lands with fee-simple tenure, the supply of freehold land is small and restricted to prime locations. Condominium units with
from 55,414 units in 1990Q1 to 249,489 units in 2009Q4. The ratio of private housing stock (including both landed and non-landed housing) to public housing stock is about 1:4, based on the cumulative stocks of private and public housing as on 2009Q1.

Currently, there are two published indices that track price movements for private non-landed housing in Singapore. One is published by the Urban Redevelopment Authority (URA), the planning agency of Singapore, and another one is developed by the Institute of Real Estate Studies (IRES) at the National University of Singapore (NUS), jointly with other government agencies and released in 2010. The URA private residential price index is a quarterly index, whereas the NUS index is a monthly index dating only from December 2005. Both are transaction-based indices constructed using a representative basket of properties weighted by transaction values of properties for each index period. The URA index uses the median price approach; the NUS index is constructed using a locally weighted regression (LWR) technique to adjust value changes of representative properties to their market values for each period.

Figure 1 shows the historical trend of the URA private non-landed residential property price index. Singapore’s private housing market has witnessed three distinct peaks in the cycles for the periods from 1975Q1 to 2010Q2. The peak in private housing demand in 1981 coincided with the Approved Residential Properties Scheme (ARPS) of the Central Provident Fund (CPF) rule, a compulsory retirement saving scheme which allows its members to draw on their compulsory retirement savings to finance purchases of private housing. Low interest rates in early-1990s – the 3-month interbank interest rate fell to 1% in 1995Q2 – led to excessive price escalation culminating in the peak in 1996. Fears of overheating in the housing market prompted the government to intervene.

---

6 The condominium stocks doubled from 55,959 units in 1998Q4 to 111,850 units in 2008. Together with apartments, non-landed housing is the largest and most popular housing form, constituting more than 71% of the private housing market by cumulative units as of 2008Q4.

7 Meese and Wallace (1991) were the first to use the nonparametric locally weighted regression technique estimate hedonic housing price indices. LWR uses a fraction of the sample to estimate a potentially highly nonlinear regression surface for each quarter.
by introducing market-cooling measures in May 1995, including reducing banks’ loan to value ratio from 90% to 80% and imposing stamp duties on buyers and sellers. The third wave of increases in housing prices started in 2004 after the economy recovered from two major shocks – the 2001 World Trade Center attack in New York and the SARS epidemic in Asia. The influx of foreign liquidity into the luxury private housing led a broad-based rebound in private housing prices. The fixed-basket URA private housing price index reached new historic heights in 2010, surpassing the benchmarked price from the 1996 peak. The increase in private housing prices spilled over to public resale housing, leading to governmental concerns over the affordability of public housing for younger Singaporean households. The government made several attempts between September 2009 and February 2011 to curb speculative activities and lower excessive price hikes in the housing market. The market cooling measures include imposing stamp duties on sellers, restricting mortgage lending, and increasing supply by stepping up the government’s land sales.\(^8\)

[Insert Figure 1]

Property purchasers are required to lodge a caveat on property transactions under the Registration of Land Title system within 2 to 3 weeks of signing an option to purchase a property. The caveat information is made available to the URA for the construction of private housing price indices. However, the basket-based URA private residential property index and its various sub-indices do not adjust for quality variations in properties as captured in transaction records.

The most commonly used procedures for adjusting for quality variations among properties are (1) hedonic price functions and (2) the repeat sales estimator. The hedonic approach directly controls for quality differences by including characteristics of the structure and location as explanatory variables for sales prices. Hedonic estimates may be biased if there are omitted

\(^8\) The government imposes stamp duties ranging from 4% to 16%, if the properties were sold in less than four years of purchase, to cream off parts of the capital gains accrued to sellers. Liquidity constraints of buyers are tightened by capping the loan-to-value ratio to 80% for the first mortgage and 60% for the second mortgage. Increasing land sales for private housing developments and stepping up public housing programs are the supply tools aimed at curbing further escalation in housing prices.
explanatory variables that are correlated with the time of sale. The repeat sales approach attempts to control for omitted variables that are constant over time by restricting the sample to properties that sell at least twice over the sample period. If the coefficients of the underlying hedonic price function and all relevant explanatory variables are constant over time, then the changes in sales price provide unbiased estimates of a quality-constant price index. However, the repeat sales price index may still be subject to bias if the characteristics of the structure change over time or if the hedonic coefficients are not constant. Moreover, by discarding sales of properties selling only once over the sample period, the repeat sales approach produces a dramatic drop in the number of observations available to construct the price index. Small sample sizes can introduce high variances, while limiting the application of the repeat sales approach to broad geographic areas for which ample sales are available.

In this study, we construct quality-adjusted price index for non-landed sub-market in Singapore using a matched sample approach. The matched sample approach controls for omitted as well as observed property characteristics by eliminating properties from the sample that differ significantly from those that sold in the base period. The matching procedure leads to much larger sample sizes than the repeat sales approach, while being less sensitive than hedonic estimation to outliers and functional form misspecification. The approach has particular advantages over conventional methods for an analysis of the Singapore housing market. Unlike in the earlier studies by McMillen (2008 and 2010) that use single-family homes in Chicago, high-rise private housing in Singapore is often quite homogeneous, with similar sets of variables representing location, amenities, and even structural attributes. In this situation, discarding single-time sales severely restricts the sample sizes. Matching produces much larger sample sizes than the repeat sales estimator. In comparison to standard hedonic price function estimation, the matching approach requires little or

\footnote{Studies by Sun, Tu and Yu (2005) and Neo, Ong, and Somerville (2006) suggest that hedonic coefficients vary significantly over time in Singapore’s private housing market.}
no pre-imposed econometric structure, while also being less sensitive to the effects of outliers. The matching approach offers relatively efficient estimates of price changes, while also reducing problems associated with non-random sample selection.

3. Quality-adjusted Price Index Estimation Methodologies

The following hedonic price function provides a general framework for constructing price indices:

$$y_{it} = \sum_{t=1}^{T} \delta_t D_{it} + \beta_t X_{it} + \lambda_t Z_{it} + u_{it}$$  \hspace{1cm} (1)

where $y_{it}$ is the natural logarithm of the sales price of house $i$ at time $t$ (t=1,…,T), $D_t$ is a variable indicating the house sold at time $t$, $X_{it}$ is a set of observed characteristics of the structure and location, and $Z_{it}$ is a set of unobserved characteristics that also influence the sales price. The price index is constructed using the vector $\delta = (\delta_1 \ldots \delta_T)'$; while other parameters may also be of interest, the objective of price index construction is to obtain an accurate measure of $\delta$.

The hedonic approach estimates equation (1) directly using the full sample of data on housing sales. Good examples of the hedonic approach include Kiel and Zabel (1997), Mark and Goldberg (1984), Palmquist (1980), and Thibodeau (1989). However, there are limitations to the hedonic approach. One limitation of the hedonic approach is that it requires a priori specification of the functional form and accurate measures of myriad variables influencing house prices. Functional form misspecification and omitted variables that are correlated with sales timing will lead to biased estimates of $\delta$.

These biases can potentially be avoided if both the explanatory variables and their coefficients are constant over time. As originally shown by Bailey et al. (1963) and subsequently extended by Case and Shiller (1987, 1989), restricting the sample to homes that sold at least twice in the sample period produces an estimating equation that provides unbiased estimates of pure time effects without having to measure $X$ or $Z$. Using equation (1), the change in the log of sales price for a home that sold at time $t$ and an earlier time $s$ is:

$$y_{it} - y_{is} = (\delta_t D_{it} - \delta_s D_{is}) + (\beta_t X_{it} - \beta_s X_{is}) + (\lambda_t Z_{it} - \lambda_s Z_{is}) + (u_{it} - u_{is})$$  \hspace{1cm} (2)
which in turn reduces to

\[ y_{it} - y_{is} = (\delta_t D_{it} - \delta_s D_{is}) + (u_{it} - u_{is}) \]  \hspace{1cm} (3)

if \( X, Z, \beta \) and \( \lambda \) are constant over time. However, the coefficient estimates may still be biased if the housing characteristics or their coefficients change over time, i.e., if \( (\beta_t' X_{it} - \beta_s' X_{is}) \) or \( (\lambda_t' Z_{it} - \lambda_s' Z_{is}) \) is correlated with \( (\delta_t D_{it} - \delta_s D_{is}) \). Unobserved renovation and concentration of sales in specific locations with high appreciation rates are examples of the omitted variables problems that cause biases in the estimation.

4. Matching

McMillen (2010) points out that the repeat sales approach can be considered to be an extreme version of a matching estimator. Matching estimators are commonly used to estimate average treatment effects. In the case of the repeat sales model, the “treatment” is whether a home sells in one period rather than another, i.e., how much would a home be worth if it sold at time \( t \) rather than in the base period? Of course, sales do not take place every period for all properties, so we are confronted with the task of constructing counterfactual values. Commonly used matching procedures would pair each home with a similar property that sold in a different time period. Averaging sales of many properties over two time periods can produce an unbiased measure of the treatment effect – the difference in sales prices across the time periods – without having to actually observe prices of an identical property across both time periods.

The following simple two-period version of the model with constant coefficients and housing characteristics illustrates the analogy between the repeat sales estimator and a matching model:

\[ y_{it} = \delta_1 + (\delta_2 - \delta_1)D_{i2} + \beta' X_i + \lambda' Z_i + u_{it}, \text{ or} \]

\[ y_{i2} - y_{i1} = (\delta_2 - \delta_1)D_{i2} + u_{i2} - u_{i1} \]  \hspace{1cm} (4)

The hedonic approach would attempt to estimate (4) directly by regressing \( y \) on \( D_2, X, \) and \( Z \). The repeat
sales model is simply a regression of $y_i - y_i$ on $D_i$ while omitting the intercept. The repeat sales version of the model is equivalent to a standard difference-in-means test: do the average log sales prices differ between period 1 and 2 for the sample of homes that sold in both periods?

Either approach provides an estimate of the expected difference in the price of a home selling in period 2 rather than period 1. This difference is written as an average treatment effect (ATE) as follows:

$$ATE = \frac{1}{n_2} \sum_{i=1}^{n_2} D_{i2} E[y_i(t_2) - y_i(t_1)],$$

which represent the expected difference in sales price between the first and second period for the sample of homes that actually sold in the second period. $ATE$ is the “average treatment effect on the treated”, where the “treated” are the $n_2$ properties that sold in the second time period.\(^\text{10}\) Since period 1 serves as the base for the price index, this expression applies directly to subsequent time periods also:

$$ATE(t_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} D_{ij} E[y_i(t_j) - y_i(t_1)] \quad (5)$$

Equation (5) is a version of a Laspeyres price index: what is the expected change in sales price relative to the base period for those properties that actually sold in time $t_j$?

The repeat sales estimator leads directly to equations (4) because the effects of $X$ and $Z$ disappear when the sample is restricted to homes with constant characteristics that have sold twice in the sample period. However, a difference in means can also be used to estimate average treatment effects when matches are not perfect. The requirements for estimating average treatment effects are less stringent than the requirements for unbiased estimates of a full set of parameters. If sales prices are appreciating at the same rate across house types and neighborhoods – an implicit assumption behind the repeat sales and hedonic approaches – then it is not necessary to have perfect matches to estimate the average treatment effect. Differences between matched observations will average out over many observations, leading to an accurate estimate of the rate of appreciation across time periods.

\(^\text{10}\) The literature on treatment effects is large and growing. Excellent overviews are presented in Ho et al. (2007) and Imbens and Wooldridge (2009).
Many approaches have been proposed for constructing matches. We use a propensity score approach: since “treatment” is a sale at time \( t \) and the “control” is a sale at time 1, the predicted values from a probit or logit model of sales time can be used to construct matches. Defining \( I_t = 1 \) if a sale occurs at time \( t \) and \( I_t = 0 \) if it sold during the base period, the propensity score is simply the set of predicted values from a probit or logit regression of \( I_t \) on the observed set of housing and location characteristics. Each observation in the base period is thus matched with an observation from the later time period that had a similar probability of sale.\(^{11}\)

In the empirical section of the paper, the data set covers the time from Jan 1995 to May 2010. We use the first quarter of 2000 as the base time. We use an algorithm that matches each observation in 2000:Q1 to similar observations in subsequent quarters. The advantage of this approach is that it assures that matched samples for later quarters are restricted to observations with similar sales in the smaller base period, which assures that the matched samples remain similar across all time periods. The following algorithm is used to construct the matched samples:

1. For each quarter \( q \) from 1995Q1 to 2010Q2 (excluding 2000Q1), estimate a logit model using all sales taking place in 2000Q1 and \( q \). The dependent variable equals one if the sale took place in quarter \( q \) and zero if the sale is from 2000Q1. The explanatory variables for the logit regressions are the same as those used for the hedonic price function estimates.

2. Use the estimated propensity score from each logit regression to match \( n_t \) observations from quarter \( q \) to sales from 2000Q1, where \( n_t \) is the number of sales in 2000Q1. Based on a random ordering of the 2000Q1 observations, each observation is matched without replacement to its closest counterpart in quarter \( q \).

At the end of this matching process, the matched-sample data set comprises approximately 62\( n_t \) observations – \( n_t \) matched sales for each quarter from 1995Q1 to 2010Q2, plus the \( n_t \) sales from the base year 2000Q1 (“approximately” because some observations may not have a close match).

After constructing the matched samples, the estimated price index is simply the series of differences between the average log-sales prices at time \( t \) and the average value in the base period,

\(^{11}\) Ho et al. (2007) present a particularly useful discussion of the practical issues involved in matching estimators. Their “MatchIt” routine in R (Ho et al., forthcoming) constructs matches very rapidly for a wide variety of algorithms. Good examples of matching estimator application in urban economics include Bondonio and Engberg (2000), Bondonio and Greenbaum (2007), Cho (2009), List et al. (2004), McMillen and McDonald (2002), O’Keefe (2004), Reed and Rogers (2003), and Romero (2009).
Wang and Zorn (1997) demonstrate a remarkable result: a period-by-period comparison of average log sales prices is exactly equivalent to the repeat sales model when applied to a data set with the same number of observations in each sample period. The repeat sales estimator controls for housing characteristics by matching each home to itself across time; the matching estimator accomplishes a similar objective by balancing the distributions of housing characteristics over time by pairing each observation with similar observations from other time periods. By eliminating extreme observations, the differences in housing characteristics average out, thereby leaving simple differences in means as an unbiased estimate of quality-constant price differences. Similar in spirit to the repeat sales estimator, period-by-period means from matched sample estimates are likely to be more efficient because sample sizes are larger when single-sale properties are included.

It is also possible to use a quantile estimator to show how the sales price distribution evolves over time. McMillen (2008) uses this approach to compare the distribution of sales prices in 1995 to 2005 and finds that the distribution shifted further to the right at higher quantiles. Note that the average predicted value at each quantile is simply the target percentile of the distribution of the dependent variable. Thus, a period-by-period series of quantiles is equivalent to period-by-period quantile regression estimates.

5. Data Analysis

For analyzing the quality effects on price changes, we use transaction data of the non-landed sub-market (which includes condominiums and apartments) from the caveat records that were made available by the URA in its Real Estate Information System (REALIS). The transactions data cover the sample period from January 1995 to May 2010. A total of 225,433 non-landed private housing sales were retrieved from the database. Based on the sales sample, we pool 129,374 transactions that came from housing units sold at least two times during the sample period, or an equivalent of
For the matching model, we match 65,556 over 61 \( n_1 \) periods during the sample periods from 1995Q1 to 2010Q2, using the 2280 housing sales in 2000Q1 as the base.

The property-level information recorded in the REALIS database includes the project name, address, floor area (\( m^2 \)), transaction price (SGD), contract date, tenure, postal district, and planning region. The island of Singapore is divided into six planning regions that include the Core Central Region, Rest of Central Region, West Region, East Region, and North Region and North-East Region. The latter two regions are combined into one due to the relatively small number of sample transactions.

Three distance variables are used to control for the accessibility of the sample of housing. Linear distances between each property and the Central Business District (CBD), the nearest mass rapid transit (MRT) or light rail transit (LRT) stations are measured. We estimate location characteristics for the sample of properties using various virtual mapping services including Streetdirectory.com. The Raffles Place MRT station is used as the reference point for the CBD. All distances are measured in straight-line kilometers. We include both the level and natural logarithm of distance to the CBD and the nearest rail station as explanatory variables to allow for some simple nonlinearity in the effect of distances on sales prices. Other amenities such as bus interchanges, popular schools, shopping malls, industrial buildings, expressways, water bodies, greenery, and public housing estate are identified, and their straight-line distances to the sample properties are measured.

In the primary school enrolment process of the Ministry of Education (MOE), children living within one kilometer of a choice school are given priority in the allocation of space in primary

---

12 Out of the 64,687 pairs of repeat sale samples, 50,215 homes were sold twice, 11,917 homes sold three times, 2,177 homes sold 4 times, 330 homes sold 5 times and the remaining 48 homes sold more than 5 times over the sample period 1995Q1-2010Q2.

13 As Singapore is a densely built-up city state connected by closely knitted transportation network, linear distances to the employment center and the stations along the mass rapid transits and the light rail transit lines are good approximations of the travel distances to the respective nodes.
Houses located near the top-ranked primary schools are, therefore, in high demand among younger Singaporean families with school-aged children. For the purpose of our analysis, we focus on 19 of the more popular primary schools, based on student performance in the primary school leaving examination results from 2005 to 2008. We include as an explanatory variable a binary variable that equals 1 for flats located within 1 km radius of at least one of the popular schools. Six other binary variables are defined to control for externalities associated with neighborhood characteristics; these include proximity to bus interchanges; shopping malls; water bodies such as a sea, reservoir, or river; greenery such as nature reserves and parks; industrial parks and buildings; expressways; and public housing estates. Proximity to these neighborhood amenities is captured by dummy variables that equal 1 if a home is within a 0.3 kilometer radius of the amenity.

Table 1 provides a list of the variables used in the empirical analysis along with some descriptive statistics. Figure 2 shows the number of observations by quarter for each of the data set. The highest sale volume was recorded in 2007 prior to the US subprime crisis. The housing market saw a quick rebound with strong sale momentum in the first two quarters of 2009. The high transaction volume between 2007 and 2010 coincided with a larger pool of repeat-sale samples during this period. The pooled repeat sale samples, however, were smaller than the matching samples in most of the period.


---

14 In Singapore, the entrance system to primary schools is based on the distance between home and school, whereas admission to secondary schools and junior colleges is determined by the scores on major examinations. Proximity to popular primary schools appears to be a more important determinant of housing prices than proximity to secondary schools and junior colleges.

15 The 1km distance is consistent with the government’s policy on the primary school entry allocation criterion. Houses located beyond 1km from the preferred school are not given the same priority in the entry selection consideration.

16 Though it might be desirable to include such building level attributes as the number of bedrooms and bathrooms, Singapore non-landed housing is quite homogenous. Area and floor level are the most important determinants of sales prices within buildings. Moreover, Large open spaces are rare given the high land costs in Singapore.
We regress the natural logarithm of sale prices on vectors of structural and neighborhood characteristics, which include controls for fixed effects for the time of sale and planning region. The regression estimates for the full-sample hedonic model and the matched sample model are summarized in Table 2. The $R^2$s range from 0.369 to 0.578 with the matched sample model with regional fixed effects explaining nearly 60% of variations in housing prices. The estimated coefficients are highly significant and the signs are consistent across the full sample and matched sample models. The coefficients on the log-distance to the CBD and to the MRT and LRT stations are positive and significant, implying premiums for access to employment centers and train stations. The relationships may not be linear, as indicated by the negative direct distance measures for the same variables. The coefficients on the neighborhood variables imply premiums to being close to amenities like public housing, industry parks, shopping centers, and bodies of water. However, homes that are close to expressways and bus interchanges may suffer from noise and heavy traffic flows. The negative effects of traffic congestion during school hours may outweigh the benefit of being close to popular schools.

The price indices for hedonic and matched sample indices are constructed using the estimated coefficients for the quarterly dummy variables, after normalizing the value of 2000Q1 to zero. The repeat sales sample is constructed by regressing the change in the natural log of sales price on a series of 61 indicator variables that equal 1 if the later sale in a repeat sales pair takes place at the time indicated by the variable, -1 if the earlier sale occurs at the time, and 0 otherwise.

Figure 3 shows the standard hedonic and repeat sale price indices along with the full sample means. The hedonic price indices track the full sample mean more closely than the repeat sale indices over the sample period. The repeat sale indices are smoother than the hedonic indices.

In Figure 4, we plot the hedonic estimates alongside two alternative approaches for estimating

---

17 The $R^2$s are reasonable despite the fact that variations in designs and interior features of sample houses are not included as explanatory variables. The fit could potentially be improved further by modelling the spatio-temporal correlation of building and spatial attributes as in Sun, Tu and Yu (2005).
price indices with the combined matched samples. The first approach for the matched samples follows directly from equation (5): once the distributions of covariates are balanced by matching sales across the control and treatment samples – 2000Q1 versus subsequent quarters – the difference in means provides a direct estimate of the rate of appreciation from 2000Q1 to quarter \( q \). Figure 4 shows that these differences in matched-sample means are remarkably close to the hedonic price index estimates. The two matched sample results are nearly identical to the base hedonic price index. The results imply that the hedonic approach is not unduly affected by outliers and unrepresentative observations in this sample of sales.

Once the observations without matches are removed from the sample, a simple comparison of means is a reliable estimator of the price index. This result is important because it suggests that other points of the sales price distribution may be estimated reliably using simple percentiles of the log of sales prices for each quarter’s matched sample. In fact, the full distribution of sales prices can be estimated using a kernel density approach. Figure 5 shows the Kernel density results for 1997 and 2007 for the full sample and the combined matched samples. The remarkable feature of Figure 5 is the much higher density at high sales prices in 2007. This result, which is consistent with McMillen’s (2008) finding for Chicago, suggests that the rate of price appreciation was much higher for high-priced properties than for low-priced homes over this period. It is important to note that this result would not have been found with standard hedonic or repeat sales procedures, whose focus on the mean carries with it the implicit assumption that the entire distribution of prices simple shifts to the right by same amount at all price levels.

Another way to see this result is to compare percentiles for the log of sales prices across time. These results are shown for the full sample in Figure 6 and for the combined matched sample in Figure 7; the lines represent the 10%, 25%, 50%, 75%, and 90% percentiles of the distribution of the natural log of sales price in the matched samples. The 10% and 90% percentiles are shown in Figure 8. Figure 8 shows large variations in price distributions for the 10% and 90% percentiles during the booms of pre-1996 and the busts in post-2007. Although prices rose markedly during the 1996 boom periods, prices dropped
less dramatically in the bust period after 2007. Overall, the variance of the sales prices distribution appears to rise significantly during boom periods.

Figures 5 and 7 illustrate the advantages of the matching approach to price index estimation. The initial matching process produces samples with similar distributions of housing characteristics. Because the distribution of characteristics is similar over time, it is appropriate to simply compare means, medians, or other points in the sales price distribution over time. Thus, the density estimates shown in Figure 5 for the matched sample show changes in the full distribution of quality-controlled sales prices over time. Figure 7 shows price indices that are analogous to standard mean-based estimates, but they provide much more information by showing how prices evolve over different parts of the distribution. Thus, the matching estimator produces a much richer set of results than is available using either the repeat sales or standard hedonic approach.

The unique composition of our sample allows us to sort homes into two groups comprising 48,231 new homes and 13,036 repeat sales. We can use the two sub-samples to conduct some robustness checks. We apply the matching estimator separately to the two housing sub-samples; the two housing indices are shown in Figure 9. The matched index for the new-house sub-sample is clearly higher than the indices for the full sample and the smaller repeat-sale sub-sample. The differences between the price indices suggest that depreciation affects the estimates, as argued in Case and Quigley (1991), and Englund, Quigley and Redfearn (1998 and 1999). The results also suggest a different interpretation for problems associated with differences between offer and reserve prices in changing market conditions, as tested in Gatzlaff and Haurin (1999): the two sub-samples show differences between market-based pricing (i.e., the repeat-sale sample) and pricing in the restricted market (i.e., developers’ pricing).

7. Conclusion

This paper demonstrates some of the advantages of treating the repeat sales approach to price index construction as a special case of a matching model. Pairing each sale in the base period to a similar sale from another period increases sample sizes compared with the restrictive repeat sales
approach of matching each property only to sales of the identical property at a different time. Recognizing the analogy between repeat sales estimation and average treatment effects makes it clear that the only difference between the repeat sales estimator and period-by-period differences in means is the repeat sales estimator places different weights on the means depending on the number of sales. However, it is not necessary to go to the extreme of matching the sale of a home only to itself to enjoy the benefits of matching estimators. Matching algorithms pair control observations with similar, untreated observations without necessarily observing the same individual in both the control and treatment samples. As long as the distributions are balanced across the control and treatment samples, differences in the covariates will tend to average out over large samples. A price index can then be constructed by simply calculating average sales prices across a series of matched samples. The matching approach also makes it easy to compare medians or other percentiles of the sales price distribution, thereby providing a more complete picture of the change in the distribution of sales prices over time.

In this paper, we have used data from Singapore for 1995-2010 to illustrate the use of a matching estimator for price index construction. The standard repeat sales approach, hedonic estimation, and the matched sample approach all indicate that the average price of a quality-constant home increased significantly during this time. Quantile estimates for the matched samples suggest that the price distribution shifted much further to the right at high sales prices during the two boom periods of 1996 and 2005-2007. The quantile results imply that the variance of sales prices increased markedly during times of rapidly appreciating sales prices – a result that would not have been found with conventional approaches for constructing price indices.

References
Bondonio, D., Greenbaum, R. 2007. Do local tax incentives affect economic growth? What mean
impacts miss in the analysis of enterprise zone policies. Regional Science and Urban Economics 37, 121-136.


McMillen, D., McDonald, J. 2002. Land values in a newly zoned city. Review of Economics and


### Table 1: Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th>Repeat Sales Sample</th>
<th>Matched Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Stdev</td>
<td>Mean</td>
</tr>
<tr>
<td>Log Sales Price (Singap)</td>
<td>13.768</td>
<td>0.551</td>
<td>13.756</td>
</tr>
<tr>
<td>Area of Unit (square m)</td>
<td>128.663</td>
<td>89.384</td>
<td>126.430</td>
</tr>
<tr>
<td>Floor Level</td>
<td>28.404</td>
<td>23.902</td>
<td>27.907</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>11.836</td>
<td>4.458</td>
<td>11.456</td>
</tr>
<tr>
<td>Log of Distance to CBD</td>
<td>2.366</td>
<td>0.512</td>
<td>2.320</td>
</tr>
<tr>
<td>Distance to MRT or LRT</td>
<td>0.486</td>
<td>0.488</td>
<td>0.493</td>
</tr>
<tr>
<td>Log of Distance to MRT or LRT</td>
<td>-1.033</td>
<td>0.708</td>
<td>-1.021</td>
</tr>
<tr>
<td>Bus Interchange</td>
<td>0.086</td>
<td>0.281</td>
<td>0.095</td>
</tr>
<tr>
<td>Popular School</td>
<td>0.461</td>
<td>0.498</td>
<td>0.472</td>
</tr>
<tr>
<td>Shopping Center</td>
<td>0.318</td>
<td>0.466</td>
<td>0.312</td>
</tr>
<tr>
<td>Body of Water</td>
<td>0.554</td>
<td>0.497</td>
<td>0.567</td>
</tr>
<tr>
<td>Green Park</td>
<td>0.533</td>
<td>0.499</td>
<td>0.543</td>
</tr>
<tr>
<td>Ind Park</td>
<td>0.406</td>
<td>0.491</td>
<td>0.397</td>
</tr>
<tr>
<td>Major Expressway</td>
<td>0.262</td>
<td>0.440</td>
<td>0.292</td>
</tr>
<tr>
<td>Public Housing</td>
<td>0.806</td>
<td>0.396</td>
<td>0.784</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>225,430</td>
<td>129,374</td>
<td>65,556</td>
</tr>
</tbody>
</table>

*Note:* All distances are measured in straight-line kilometers.
Table 2: Hedonic and Matched Sample Regressions

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th></th>
<th>Matched Sample</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Estimate</td>
<td>Estimate</td>
<td>Estimate</td>
</tr>
<tr>
<td>Area of Unit</td>
<td>0.003***</td>
<td>0.002***</td>
<td>0.003***</td>
<td>0.003***</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Floor Level</td>
<td>-0.001***</td>
<td>0.000***</td>
<td>0.013***</td>
<td>0.008***</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>-0.047***</td>
<td>0.002</td>
<td>-0.055***</td>
<td>-0.010***</td>
</tr>
<tr>
<td></td>
<td>0.002</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>Log of Distance to CBD</td>
<td>0.420***</td>
<td>-0.014</td>
<td>0.459***</td>
<td>0.059***</td>
</tr>
<tr>
<td></td>
<td>0.014</td>
<td>0.012</td>
<td>0.015</td>
<td>0.014</td>
</tr>
<tr>
<td>Distance to Station</td>
<td>-0.170***</td>
<td>-0.046***</td>
<td>-0.166***</td>
<td>-0.095***</td>
</tr>
<tr>
<td></td>
<td>0.015</td>
<td>0.013</td>
<td>0.016</td>
<td>0.014</td>
</tr>
<tr>
<td>Log of Distance to Station</td>
<td>0.165***</td>
<td>0.141***</td>
<td>0.126***</td>
<td>0.146***</td>
</tr>
<tr>
<td></td>
<td>0.012</td>
<td>0.010</td>
<td>0.013</td>
<td>0.012</td>
</tr>
<tr>
<td>Bus Interchange, .3 km</td>
<td>-0.115***</td>
<td>-0.014*</td>
<td>-0.043***</td>
<td>0.060***</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.006</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>School, 1 km</td>
<td>-0.087***</td>
<td>-0.068***</td>
<td>-0.074***</td>
<td>-0.045***</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.004</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>Shopping Center, 0.3 km</td>
<td>0.169***</td>
<td>0.061***</td>
<td>0.148***</td>
<td>0.025***</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.004</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>Water, 0.3 km</td>
<td>0.154***</td>
<td>0.050***</td>
<td>0.284***</td>
<td>0.169***</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.006</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>Park, 0.3 km</td>
<td>0.012</td>
<td>-0.038***</td>
<td>-0.077***</td>
<td>-0.099***</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.006</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>Industrial Park, 0.3 km</td>
<td>0.107***</td>
<td>0.109***</td>
<td>0.093***</td>
<td>0.089***</td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.006</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>Expressway, 0.3 km</td>
<td>-0.034***</td>
<td>-0.135***</td>
<td>-0.061***</td>
<td>-0.127***</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.004</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>Public Housing, 0.3 km</td>
<td>0.054***</td>
<td>0.167***</td>
<td>0.047***</td>
<td>0.130***</td>
</tr>
<tr>
<td></td>
<td>0.006</td>
<td>0.005</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>Yr-Qtr Fixed Effect</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Planning Region Fixed Effect</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>R-square</td>
<td>0.369</td>
<td>0.548</td>
<td>0.424</td>
<td>0.578</td>
</tr>
<tr>
<td># obs</td>
<td>111,072</td>
<td>111,072</td>
<td>64,330</td>
<td>64,330</td>
</tr>
</tbody>
</table>
Figure 1: Price Index for Non-Landed Residential Properties and interest rates
Figure 2: Number of Observations by Quarter
Figure 3: Standard Price Indices
Figure 4: Hedonic Price Index and Matched Sample Means
Figure 5: Sales Price Densities
Figure 6: Price Quantiles for the Full Sample
Figure 7: Price Quantiles for the Matched Samples
Figure 8: The 10th and 90th Percentiles of Log-Sale Prices
Figure 9: Matched Indices based on New Sale and Repeat Sale Samples