A New Approach for Constructing Home Price Indices in China:

The Pseudo Repeat Sales Model

Xiaoyang GUO[1,2], Siqi ZHENG[1,*], David GELTNER[2] and Hongyu LIU[1]
(1: Hang Lung Center for Real Estate, Tsinghua University;
2: Center for Real Estate, Massachusetts Institute of Technology;
* Contact author, zhengsiqi@tsinghua.edu.cn)

Abstract:

Due to data and methodology constraints, there is a lack of good quality-controlled residential price indices publicly available in China. New home sales account for quite a large share of total home sales (87% in 2010) in Chinese cities, As a result, the standard repeat sales approach cannot be employed, as a new housing unit only appears once on the market. The hedonic method may be more suitable in principle, but it is vulnerable to an omitted variables problem which may be more significant in Chinese cities due to extremely dynamic urban spatial structure development and fast infrastructure construction.

Taking advantage of a unique feature of residential development in Chinese cities, we develop a “pseudo repeat sale” model (ps-RS) to construct more reliable quality-controlled price indices for newly-constructed homes. The new homes are developed in the form of residential complexes. Each complex is developed by a single developer and contains a number of high-rise residential buildings. Each housing unit within the same complex shares the same location and community attributes, as well as similar physical characteristics (such as structure type, architecture style, housing age, etc). Of course, there may still be important differences in unit size, number of bedrooms, floor level within the high-rise, and the direction the main bedroom faces. Based on specific criteria, we match two very similar new sales within a complex to create a “pseudo-pair”. We are able to generate a vast number of such pairs, many more than in traditional repeat sales models. By regressing the price differential onto the within-pair differentials in unit-specific physical attributes as well as the usual repeat-sales time dummy variables corresponding to the index periods (locational and community variables are cancelled out), we are able to construct a ps-RS price index for new homes.

This ps-RS price index approach not only addresses the problem of lack of repeat-sales data and the omitted variables problem in the hedonic, but also addresses the traditional problems with the classical repeat-sales model in terms of small sample sizes or sample selection bias. Another advantage of this index is its transparency and ease of understandability for better communication with non-specialized constituencies (government and private sector policy makers, investors, and analysts).

We test the approach using a large-scale micro transaction data set of new home sales from January 2005 to June 2011(469,070 observations) in Chengdu, Sichuan Province. We estimate our ps-RS indices and compare them with a corresponding standard hedonic index. The two indexes show similar overall price appreciation patterns, but the ps-RS index has less volatility and larger first-order autocorrelation than the hedonic index, suggesting that the ps-RS exhibits less random estimation error.

The ps-RS approach may be suitable for any rapidly urbanizing country in which new home sales dominate the housing market and where the new housing stock is constructed in large-scale complexes consisting of many relatively homogeneous individual units.

Keywords: Residential Price index; repeat sale; hedonic; pseudo repeat sale index, matching, rapid urbanization

1. Introduction

In the world of transaction price indices used to track market dynamics in housing, the problem of controlling for heterogeneity in the homes transacting in different periods of time is perhaps the most crucial challenge. The simple mean or median values of sale prices per square meter are not reliable because the size, quality, and components of the homes being sold keep changing over time. The two major methods in the academic literature for addressing this challenge are the hedonic and repeat sales approaches. Of these two, in the U.S., only the repeat-sales approach has seen widespread regular production and publication in official or industry statistics (for example, in the OFHEO and S&P/Case-Shiller home price indices).

Kain and Quigley (1970) decompose the components of housing price dynamics using the hedonic model, from which a housing price index is generated by controlling for home transactions′ physical and location attributes. Other pioneers of hedonic price modeling were Court (1939), Griliches (1961), and Rosen (1974). Two specific methods are proposed to construct the hedonic housing price index. The first method assumes constant relative preferences for housing attributes over time, and estimates a single hedonic regression for the whole historical sample (pooled database), using time-dummies to capture the price evolution over time, and constructing the price index from the coefficients of the time dummies. The second method is to run separate hedonic regressions for each period, and construct the price index as the predicted value from each period′s regression model of a standard (or "representative") housing unit that is held constant across time.

The repeat sales model was introduced first by Bailey *et al* (1963) to calculate a housing price change indicator using only properties that sold twice or more in the historical sample. The basic idea is to regress the percentage (or log) price *changes* between consecutive sales of the *same* properties onto a right-hand-side data matrix that consists purely of time-dummy variables corresponding to the historical periods in the price index. The time-dummies assume a value of zero before the first sale and after the second sale. The model was largely ignored for two decades before being independently "rediscovered" (and enhanced) by Case and Shiller (1987, 1989).

The repeat sales model has some advantages and disadvantages from an econometric perspective, as will be reviewed shortly. But before delving into the econometrics, we should note that one advantage of the repeat sales model that is beyond the technical academic perspective is its relative simplicity. This may partially account for why it has been used much more than the hedonic model in actual practice in industry and government. The repeat sales model is relatively easy for a less technical, non-specialized constituency to understand and feel comfortable with. It is easy for users to understand a meaningful price-change metric as that of, and within, the same property between consecutive "buy" and "sell" transactions, in which the same owner or investor is on both ends of the round-trip investment experience. However interesting the *cause* of the price change (e.g., whether it is due to the opening of a new subway station or a new school, as can be studied through hedonic modeling), the result is the same in terms of asset price and value impact for the property investor/owner. The repeat sales model trades off an ability to more deeply analyze the cause of price changes from an urban economics perspective, for a more parsimonious specification that has less challenging data requirements, is more readily understandable by non-specialists, and leaves less room for debate about exactly what is the "correct" or "best" model specification.

From an econometric perspective, as demonstrated by Clapp and Giacotto (1992), the repeat sales model is mathematically equivalent to the pooled-database hedonic model, as it is the differential transformation of the hedonic model, assuming that the coefficients of the attributes are constant. In other words, subtract the hedonic model of the second-sale price from the hedonic model of the first-sale price, and the hedonic variables cancel out (assuming they're constant with constant coefficients), and what's left is just the time-dummy variables as specified in the repeat-sales model.

Potentially different results from the two models then come only from the difference in the sample selection of the estimation database, with only properties having sold more than once able to be included in the repeat-sales model′s sample. Therefore, the repeat sales model can be treated as a special estimation sample case of the pooled-database hedonic.[1]

In spite of the popularity of both models, the discussion about their shortcomings has never stopped in the urban economics and econometrics literature. The hedonic model is perhaps superior in theory, but often weaker in practice, because of the omitted variables problem in real world datasets. As a result, it has been claimed that all hedonic based housing price indices are more or less biased (Quigley, 1995). The parsimony of the repeat sales model, on the other hand, probably tends to make it more robust to omitted variables. But its weakness is the limited sample size and sample selection bias, because of its need for repeat-sales. Sample selection bias or small sample sizes can be addressed in various ways, but these remain concerns in the classical repeat-sales index (Meese and Wallace, 1997; Gatzlaff and Haurin, 1998).[2]

A number of methods have been proposed to address these issues. Case and Quigley (1991) developed a hybrid model to combine the advantages, and avoid the weaknesses, of the hedonic and repeat sales models. Quigley (1994) simplified the hybrid model by adding a stochastic effect estimated from the repeat-sale sample, representing the real average price change over time. Case, Pollakowski and Wachter (1991) empirically tested and compared three groups of housing price indices models, finding that the hybrid model appeared to be empirically more efficient than either the

[1] It should be noted that while the RS model *can* be derived as the differential of the pooled-database hedonic model, it need not be so derived. The RS model can stand on its own as a primal specification. As such, the only assumption is that the time-dummy coefficients represent *all* of the longitudinal change in pricing, from whatever source or cause, between the first and second sales. Viewed from the hedonic perspective, such price changes may reflect changes in hedonic coefficients (changes in implicit prices of the hedonic attributes), changes in the values of the hedonic attributes (which presumably is minimal within the same unit), or movement in an “intercept” in the hedonic specification (which might reflect general market conditions, relative balance between supply and demand).

[2] It should also be noted that repeat-sales sample sizes may not necessarily be much if any smaller than hedonic sample sizes once one considers the need for all of the hedonic observations to include good values for a range of hedonic variables, whereas the repeat-sales model needs only the sale price and date.

hedonic or repeat sales model, and that the difference between the results of the hedonic and hybrid comes from the systematic differences between single transactions and repeat transactions. Similar results have been verified by a large literature (Englund, Quigley and Redfearn, 1999; Hansen, 2009).

An interesting perspective to take on the repeat sales model is to view it as one (extreme) solution to a matching problem. The objective is to match or pair sales observations together according to certain specific criteria so as to drop out unobservable attributes, making the model more parsimonious and less in need of lots of good hedonic data. In the classical repeat sales model, the matching criterion is extreme in that a sale is matched only to its previous sale of the exact same property, so that as much as possible of the variation in location and physical attributes are cancelled out (except for property age and possibly some innovations in the neighborhood or improvements in the house). McMillen (2010) suggests a more open matching approach as an alternative. Deng, McMillen and Sing (2011) specially expanded this approach and applied it to Singapore's residential market. They predict the sale probabilities of all the transactions, and then respectively match each transaction after the base period with a transaction in the base period with the closest sale probability. This matching approach preserves a larger sample than the classical repeat sales model while requiring less variables and form specification than a classical hedonic model. However, it is complex and may be difficult for non-specialists to understand, and it may have substantial data requirements to estimate the sales probability model which is required in the method. While Singapore's residential market shares some common features with that in China, the Singapore market lacks some of the extreme challenges found in Chinese cities.

Consider some of the unique features in China's urban residential market. New home sales account for an exceptionally large share of total sales (87% in 2010) due to a growth rate in the Chinese economy and urbanization that is truly unprecedented in world history. The classical repeat sales approach cannot be employed because a new housing unit only appears once on the market. Yet the hedonic method may face more than its usual challenges because the omitted variables problem may be more severe in Chinese cities due to such rapid urban spatial structural evolution and fast infrastructure construction.

The proposal in this paper is to develop a new type of “repeat sales” model, which we dub “pseudo repeat sales” (ps-RS). In effect, we propose a new matching criterion that is particularly appropriate in Chinese cities. We deal with the omitted variables issue by employing a within-complex matching criterion instead of the more stringent, classical same-unit criterion[3]. This approach not only addresses the problem of lack of repeat-sales data and problematical hedonic variables observation, but also addresses the traditional problems with the classical repeat-sales model in terms of small sample sizes or sample selection bias. More specifically, the proposed model is (in fact must be) a hybrid repeat sales/hedonic model (because the units are not identical) of the type noted previously has been demonstrated to have desirable features in the econometric literature. But the hybrid (hedonic) component of the model is small and relatively easy to understand and for which good data can be easily obtained, thus retaining the essential “repeat sales” characteristic of the model. We will argue that the result is a more reliable housing price index more suitable for the new residential markets in Chinese cities.

The rest of this paper is organized as follows: Section Two describes the features of the new-home market in Chinese cities and how those features affect the choice of housing price index construction methodology. We introduce our strategies for developing the ps-RS index in Section Three. After data description in Section Four, the index calculation results are illustrated in Section Five, followed by a quantitative comparison of the ps-RS with the standard hedonic method (which is the only realistic alternative since classical repeat sales is not possible). Section Six concludes.

[3] The matching criterion can also be limited to sales only within the same buildings. However, this reduces the pseudo-sample size, and exploratory analysis indicated that the within-complex criterion works better in Chengdu.

2. Features in China's Urban Housing Market and Their Implications for Price Index Construction

Before the 1980s, urban housing in China was allocated to urban residents as a welfare good by their employer (the work unit) through the central planning system. Workers enjoyed different levels of housing welfare according to their office ranking, occupational status, working experience and other merits. Governments and work units were responsible for housing construction and residential land was allocated through central planning (Zheng et al, 2006). Since the 1980s, most of the work-unit housing units have been privatized. By the end of the 1990s, housing procurement by work units for their employees had officially ended and new homes would be built and sold in the market (Fu et al, 2000). Developable land was supplied and regulated by the government through long-term leases. The real estate market took off, and massive land development took place in many Chinese cities. Sales of newly built residential properties reached 933 million square meters in 2010, with an average annual growth rate of about 20% in the last 10 years.[4]

With the fastest urbanization in world history (almost 500 million people urbanized from 1980 to 2010), massive investment in urban transport infrastructure, and the rapid growth of the service sector in Chinese cities since the beginning of the 1990s, a more specialized land-use pattern has emerged. We see that the central business district (CBD) has greatly expanded while residential land use has extended into suburbs. Industrial land use has been pushed out from the center towards outlying urban locations. Urban built-up areas have quickly expanded and new mass housing complexes have been largely built around the fast expanding urban fringes. This dynamic evolution of urban form brings a big challenge in constructing home price indices using the hedonic method. Given the data availability constraints it is difficult

[4] To put this in some perspective, the peak year of housing construction in the U.S., 2005, saw less than 300 million square meters built (in houses that were on average more than twice the size of housing units in China).

to fully quantify or control for location attributes (even if the exact address is known). For instance, failing to fully control for the suburbanization trend will lead to a downward biased index as more distant locations sell at a discount (other things equal). And as physical quality of housing units and of the complexes in which they are developed has greatly improved with the rapid rise in per capita incomes, it becomes more important and more difficult than in more mature economies for hedonic variables to fully reflect the quality improvements.

On the other hand, the secondary (resale) market for existing homes has been slow to develop. The poor marketability of the old housing stock was reflected by the low turnover of existing homes relative to new home sales in Chinese cities. One reason was deficient private property rights in privatized work-unit-provided dwelling units—the owner-occupants' legal title to their homes was ambiguous and not fully marketable. In addition, resale market institutions, including real estate listing services, title transfer and brokerage were still under development (Zheng et al, 2006). According to the National Statistics Bureau, 87% of the total housing sales came from the newly-built housing market in 2010. The standard repeat sales method is of course not able to construct home price indices for this dominant component of the Chinese housing market, because each unit only transacts once.

An important feature in the new housing market is that new housing is supplied by real estate developers in the form of large-size residential complexes. A typical residential complex developed by a single developer usually consists of a dozen of high-rise condominium buildings that share the same location attributes, common architectural design, structure type and community/property services. There are very small within-complex differences among housing units such as floor number (height above the ground within the building), unit size, number of bedrooms, and the direction the main bedroom faces. This unique feature gives us an opportunity to develop a "pseudo repeat sales" model (ps-RS).

In the ps-RS method we match two very similar new sales within a complex or even within a building, creating a paired sale observation. We call these pairs pseudo repeat sales because the two units are not exactly the same unit. Rather, they are quite similar, much more so than different individual houses typically are in most U.S. developments.[5] But the approach is essentially like the classical repeat sales model in that we regress the price differential between the first and second sale onto time-dummy variables representing the historical periods of the price index using the same specification as classical repeat sales models. In addition, however, because the units are not exactly the same, we must incorporate some elements of the "hybrid" form of price index model that includes elements of both the hedonic and repeat sales models. Thus, in addition to the standard time-dummies, the regression's independent variables include indicators of the relatively small and easy to measure within-pair differentials in physical attributes between the two units (such as number of bedrooms and floor number).. But the major and most problematical hedonic variables, the locational and community variables, cancelled out of the model just as they do in the classical repeat sales specification. In this way we are able to mitigate the omitted variables and data problems that plague the hedonic approach.

3. Index Construction Methodology

3.1 Matching Process

The standard repeat sale model can be regarded essentially as a specific matching approach. Its matching space is the same house, which means that only repeated transactions of the same house can be matched into pairs. This extremely small matching space implicitly restricts the matching rule to be the same location and physical attributes (except for age and possible renovation). Here the matching space

[5] At least since the days of Levittown shortly after World War II. However, some U.S. housing developments even today (or when/if that industry ever gets back on its feet) are characterized by fairly homogeneous houses, and in fact the ps-RS technique might be a way worth exploring to build an interesting index of U.S. hew home price evolution.

is defined as one house.[6]

In our pseudo repeat sale model, we expand the matching space from one house to a residential complex. As mentioned above, all housing units in the complex share the same location and neighborhood attributes, and a subset of physical attributes. Applying within-pair first difference will cancel out all the above variables, including both observable and unobservable ones. Those physical attributes with within-pair differences will be left on the right-hand side as independent variables reflecting the "hybrid" specification of repeat sales and hedonic modeling.

Index frequency along time horizon should be chosen before doing the matching work. Given the rich transaction data set in Chengdu, we estimate a monthly price index.

The rule to generate a pair is to match one transaction with its most temporally adjacent transaction in the same complex. Here is a simplified example. Suppose we have four periods in total. Complex A has 3 transactions in the 1st period, 2 transactions in the 2nd period, and zero transaction in the 3rd period, and 3 transactions in the 4th period (Figure 1). When we consider the 3 transactions in the 1st period, their most adjacent transactions are the 2 observations in the 2nd period. Thus 6 pairs will be generated (2x3=6). Since no transaction in the 3rd period, when stand at the 2nd period and look forward, the 4th period is the most adjacent period. Another 6 pairs will be generated by these two periods. So our matching rule yields 12 pairs altogether.

*** Insert Figure 1 about here ***

We do not match the transactions in the 1st period and those in the 4th period into pairs because they are not "adjacent" transactions. The rationale behind is that including "non-adjacent" transaction pairs will cause redundancy (and larger colinearity for the

[6] Age per se is not something that should be controlled for if the focus of the index is to track the price change experienced by the homebuyers (investors). Buildings, like people, cannot help but age (alas).

time dummies) – the price change between the 1st and 4th periods is the linear combination of the price change between the 1st and the 2nd periods plus that between the 2nd and the 4th periods .

Though Complex A has no transaction in the 3rd period, another complex may have some transactions in that period. Since the whole sample consists of thousands of complexes, every period will be included in the new sample of transaction pairs on which our ps-RS is regressed.

3.2 Regression Model

The standard hedonic model to construct a housing price index is shown as Equation (1) (Quigley, 1991), where P_i is the house sale *i*'s total transaction value, $X_{k,i}$ is its k^{th} physical or location attribute at least some of which may be invariant over time, $D_{j,i}$ is the time dummy which equals 1 if the sale occurs in period *j*, otherwise equals 0, and ε_i is the error term.

$$\ln P_i = \sum_{k=1}^{K} \alpha_k \ln X_{k,i} + \sum_{t=1}^{T} \beta_t D_{t,i} + \varepsilon_i \qquad (1)$$

Now we turn to our pseudo repeat sale model. Here complexes are indexed by *j*, periods (months) are indexed by *t*. Within complex *j*, house *a* in month *r* and house *b* in month *s* are adjacent transactions (*s*>*r*), and the two make a matched pair. Based on equation (1), a differential hedonic regression (ps-RS model) is expressed as Equation (2). D_t is the time dummy representing the time the sale occurs. D_t=1 if the later sale in the pair happened in the month *t=s*, D_t=-1 if the former sale in the pair happened in month *t*=*r*, and D_t=0 otherwise.

$$\ln P_{b,s,j} - \ln P_{a,r,j} = \sum_{k=1}^{m} \alpha_k (\ln X_{b,s,j,k} - \ln X_{a,r,j,k}) + \sum_{t=1}^{T} \beta_t D_t + \varepsilon_{s,r,b,a,j} \quad \mathbf{(2)}$$

It is clear that our ps-RS model also follows the assumption in the classical repeat sales model, which assumes that any change over time in pricing is captured in the time-dummy coefficients[7].

3.3 Weighting Adjustment

In Equation (2) the observation is a pair. A potential problem is that by generating those pairs, the original sample size distributions over time and across complexes will be changed, relatively speaking. Consider two adjacent periods *r* and *s*, and suppose there are N_r and N_s observations in these two periods respectively. In the standard hedonic model the number of observations will be ($N_r + N_s$), while this number will increase to ($N_r * N_s$) in our pseudo repeat sale model. If N_r and N_s are big numbers, this amplification effect will be significant and bring in estimation bias to the OLS regression. This is also true across complexes. A large complex will generate an even much larger set of pairs than a small complex.

Weighted OLS is introduced to return the weight of each observation in the ps-RS model back to its original weight in a standard pooled-database hedonic model. Specifically, for the pairs of month *r* and *s* in complex *j*, the weight is:

$$w_{r,s,j} = (N_{r,j} + N_{s,j}) / (N_{r,j} \cdot N_{s,j}) \qquad \textbf{(3)}$$

4. Index Estimation and Discussion

4.1 Data

A large-scale transaction database is used to construct the ps-RS Index. The database contains the full records of Chengdu’s new residential sales from January 2006 through December 2011, consisting of 2152 complexes and altogether 469,070

[7] In the classical RS specification, where the hedonic variables are dropped out, we need not necessarily derive the RS model from the constant-attributes (pooled database) hedonic model. The price changes picked up in the RS model time-dummy coefficients may reflect changes in implicit prices, or they may reflect a movement in some sort of “Intercept” in the hedonic model. (For instance, the time dummies in a classical same-house RS model also reflect the aging of the house). So the RS model does not require that we assume constant implicit prices.

housing units after data cleaning. The information in the database includes each transaction's total purchase value, physical attributes (unit size, unit floor number, building height in floors, the number of rooms, etc.), and location attributes (the distance to the city center, and zone ID there being 33 zones[8] defined by the Chengdu Local Housing Authority). Table 1 shows the descriptive statistics of these variables.

*** Insert Table 1 about here ***

4.2 Index Estimation Using ps-RS Model

The matching space is defined as the complex in our ps-RS model. 22 million pairs are generated from the 469 thousand transactions in 2152 complexes. Table 2 shows the descriptive statistics of those pairs.

*** Insert Table 2 about here ***

Equation (2) is regressed over all the pairs using WLS, with standard errors clustered by complex. Table 3 represents the estimated results. All the coefficients of the physical attributes are statistically significant and have the expected signs. This model can explain 87.4% of cross-pair differences in price growth. Based on the coefficients of the time dummies, the ps-RS Index is calculated as the black solid line shown in Figure 2.

*** Insert Table 3 about here ***

*** Insert Figure 2 about here ***

[8] We divide the urban space of Chengdu into 33 zones by two rules: the ring-road and the direction. Chengdu is a monocentric city ,there are four main ring-roads, including the inner ring-road in the most inside of the city and another three ring-roads successively from inside to outside named as the 1st, the 2nd and the 3rd rong road. The four ring roads divide the urban space into five concentric ring areas with different distances to the city center. On the other hand, in terms of spatial direction, the urban space can be grouped into North, Northeast, East, Southeast, South, Southwest, West, Northwest and the Center. Spatially, the Center area is completely overlapped with the area inside the inner ring-road, and all the other 4 concentric areas divided by the ring-roads are further separated into 8 zones for each by the directions. As the result, we have 1 center zone and other 32 surrounding zones, with about 18.6 square kilometers for each zone on average.

The other two index lines in Figure 2 are for comparison. The short-dashed line is a median price index which is a simple series of median value of sale price per square meters in every month. The long-dashed line is a hedonic price index calculated based on the hedonic regression shown in Table 4. The three indices in Figure 2 intertwine with each other and have a similar overall trend and similar turning points. Before mid-2007, all three indices move along the same path. After a short shoot up in later 2007, the market dropped down in 2008 during the worldwide financial crisis. From the beginning of 2009, thanks to stimulus policies against the crisis such as expanded credit availability and huge government direct investment, the market turned up rapidly and kept rising until early 2011 when tight regulations were implemented. After that, the market has kept stagnant with a flat price trend. Thus, all the indices tell a similar story that conforms well with general qualitative knowledge of the market.

In most periods, the ps-RS index and hedonic index are lower than median price index. Ps-RS Index shows less volatility than the other two indices. It is easy to understand that the median price index is not as accurate because it is not quality controlled and does not employ econometric optimization to minimize noise. It seems that housing quality is rising over time, so the median price index tends to over-estimate the price growth. In the next sub-section we will focus on explaining the observed gap between the ps-RS index and the hedonic index. Then we turn to evaluate the quality of the two indices.

*** Insert Table 4 about here ***

4.3 Understanding the Gap between ps-RS Index and Hedonic Index

Recognizing that the ps-RS model derives from the differentials of the hedonic model within the matched pairs is helpful to understand the gap between two the index series. The two transactions in a pair are quite homogeneous in physical and location attributes. The small within-pair differences in physical characteristics, such as floor number, unit size and number of rooms, are also easy to be identified. The ps-RS regression will drop out all observed and unobserved common attributes, thus we are able to mitigate the problem of omitted variables or variables that are difficult to measure reliably and obtain index results that are more robust. This is the advantage of the parsimony in the RS specification. In contrast, omitted variables or hedonic variables that are difficult to properly value or quantify can bias an index estimated from a hedonic model. Therefore, the gap between the hedonic model and the ps-RS model can be attributed to the unobserved variables cancelled out in the ps-RS model. Indeed, since the ps-RS model uses *all* transactions (unlike the classical RS model that can only use repeat-sales), it would seem that omitted (or mis-valued) hedonic variables must be the major source of any difference between the hedonic and the ps-RS indices in Figure 2. This is different from a typical comparison between hedonic and classical RS indices, as in that case the estimation databases are also different.

There are also two other sources of difference between the ps-RS and hedonic indices, which may partly explain the difference we observe in Figure 2. While the ps-RS model is based on all and only the same transactions as the hedonic model, the matching process generates a *much* larger (pseudo) sample size for the ps-RS model than what the hedonic model has to work with. This larger sample size should help the ps-RS model to be estimated more precisely, resulting in less noise in the index. Finally, a third source of difference between the two indices could arise from the use of the differential specification in the ps-RS model versus the undifferenced (levels) specification in the hedonic model. The ps-RS model directly estimates longitudinal price *changes*, whereas the hedonic model directly estimates price levels as of one point in time (and the hedonic index of longitudinal price changes is then only constructed later from the differences in the hedonic model's time-dummy coefficients). The longitudinal differencing in the underlying ps-RS regression model may affect the results.

Returning to what we believe is the major source of difference between the ps-RS and hedonic indices, suppose the problematical omitted variable in the hedonic index is a positive attribute favored by households and the share of transactions with this attribute is rising over time. For example, suppose newer housing units built more recently have higher quality of the finishes on the flooring, walls and ceilings, or maybe higher quality of the heating and air conditioning systems, air and water filtration systems, or better kitchen/bathroom appliances, but the hedonic database does not have any information about quality improvement except of the number of rooms. Then the hedonic index will tend to overestimate the rate of price growth. It will in effect attribute the value of higher physical quality of housing units to the housing market condition (when in fact these represent the market for better physical quality of apartments). In such a case we would see the ps-RS index tending to track below the hedonic index. In reality, with such rapidly rising per capita income, it would seem likely that the new housing units have been incorporating more and more favorable attributes in terms of the physical characteristics within the units.

If, on the other hand, the share of the transactions with such favored (omitted) attributes is declining, then the ps-RS index will track above the hedonic index. In the case of negative (unfavorable) attributes, the situation is the opposite to what we have just described. One typical case is that the rapid urbanization has meant that location attributes may be inevitably tending to be less favorable (farther away from the CBD, although mitigated perhaps by transport infrastructure improvements and rising automobile ownership). It is possible that not all of these changes can be completely captured or accurately measured in the hedonic attributes database.

To give an intuitive example, we run two hedonic regressions, one with the distance to the city center variable (*D_CBD*) and the other without it. Of course, we cannot show an actual example of an actual omitted variables effect because by definition we don't have data on actual omitted variables. So, for this example, imagine that we somehow couldn't actually get accurate CBD distance data for the hedonic model. The suburbanization trend of residential development in Chengdu is quite significant. Suburban areas are under-developed with less infrastructure and consumer amenities, and of course, commuting to the CBD takes longer and/or is more expensive. Thus, *D_CBD* is a negative attribute (it has a negative sign in the hedonic equation). Figure 3 shows how the average distance to the city center by month of unit sale rises over time in our sample. Figure 4 shows the two hedonic indices. It is obvious that the line without controlling for *D_CBD* under-estimates price growth. If we couldn't actually accurately measure distance to the CBD, i.e., if *D_CBD* were an omitted variable, then Figure 4 shows how much we would erroneously attribute the effect of CBD distance to the price trend over time instead of to that omitted variable.

*** Insert Figure 3 about here ***

*** Insert Figure 4 about here ***

Now consider that, while *D_CBD* is observed, it may not be a sufficient variable to fully capture the suburbanization effect. If significant indicators of suburbanization are omitted, the hedonic index in Figure 2 will be biased downward. In Figure 2 we can see the hedonic index and the ps-RS index are intertwining with each other, which means that besides the suburbanization effect, some other omitted variables are playing a role. Recall that in our hedonic model we only have two location variables (*D_CBD* and zone ID) due to data availability constraints. The zones are relatively big – 18.6 square kilometers on average. We are missing within-zone location attributes, either positive or negative. Our ps-RS index helps us to implicitly control for those unobservables.

With the above in mind, it is nevertheless interesting that there is in fact not much difference in the overall price trend between the hedonic and ps-RS indices in Figure 2. This may suggest that omitted variables are not a great problem. Or, it may mean that different omitted variables offset each other and cancel out, in the case of Chengdu over the particular historical period covered. In any case, the ps-RS model is more robust to these types of problems in principle. Thus, it is the ps-RS model that is providing a sort of "check" on the hedonic model. In this case, the hedonic model seems to perform about as well as the ps-RS model in terms of overall trend. But the ps-RS model is easier to construct in terms of its data requirements. And, as noted at the outset of this paper, the ps-RS model may have additional practical advantages over the hedonic approach in terms of ease of understandability or communication to practitioners and policy makers.

To provide more background information, here we also compare our ps-RS index with the official housing price index released by the National Bureau of Statistics of China (NBSC) (so called "70-index" for 70 Chinese cities). Figure 5 shows the two indices for Chengdu from 2009M3 to 2010M12 (we are only able to find systematic NBSC index series for this period). NBSC index was calculated by simply averaging developers' self-reported price changes comparing to previous month. It is believed that developers always cheated on this by reporting much lower price changes than what was really happening, so the credibility of this NBSC index has long been criticized. We can see that in Figure 5 the NSBC index is significantly lower than our ps-RS index (of course also much lower than the hedonic index).

4.4 Comparing Indices Regarding Random Error

In this section we adopt two useful statistics to compare the ps-RS index, hedonic index and median price index, in terms of random statistical estimation error, the type of error that can impart "noise" into the index.[9] Geltner and Pollakowski (2008, as

[9] With large transaction samples such as available in typical Chinese cities, purely random error may not be a major problem, as it is due to statistical estimation error which is typically a problem with small sample sizes. Of greater concern may be sources of index bias, as we have discussed in the preceding sections. However, even with large datasets it is still desirable to minimize random error, as noise can obfuscate the "signal" or information contained in the index returns, and make the index less useful.

reported in Bokhari and Geltner, 2010) describe a model of index noise which suggests two indicators will generally be useful to quantify a comparison of the relative amount of noise in two or more indices: the volatility and the first-order autocorrelation (AC(1)) in the index returns. Traditional econometric measures based on the underlying regression, such as standard errors and signal/noise ratios, are not as appropriate for judging price indices because they are based on the residuals from the regression models underlying the index. Yet these residuals do not really measure the accuracy of the index returns. In theory an index could be perfectly accurate, exactly measuring the true market average return each period, yet the regression model would still have residuals and the index coefficients might still have large standard errors, resulting simply from the dispersion of individual property prices around the market average. The index volatility and AC(1) directly reflect the accuracy of the index returns. Other things being equal, the lower the volatility and the higher the AC(1), the more accurate (less noisy) is the index.

Label the *true* return of the market housing price in period *t* as r_t (measured as the log price difference). The returns are arithmetically added across time to build the true market value level, M_t, (in logs) as equation (4). On the other hand, label the index as of the end of period *t* as I_t, in equation (5).

$$M_t = M_{t-1} + r_t \qquad \textbf{(4)}$$

$$I_t = M_t + \varepsilon_t \qquad \textbf{(5)}$$

The ε_t term is the index-level random error, the error that causes noise and therefore matters from the perspective of index users. Noise can be modeled as having zero mean and no correlation with anything else. It is important to note that noise does not accumulate over time. For an index beginning *T* periods ago, we have:

$$I_t = M_t + \varepsilon_t = \sum\nolimits_{t-T-1}^{t} r_i + \varepsilon_t \qquad \textbf{(6)}$$

From equation (6), we obtain a formula for noise in the index return:

$$r_t^* = I_t - I_{t-1} = r_t + (\varepsilon_t - \varepsilon_{t-1}) = r_t + \eta_t \qquad \textbf{(7)}$$

Where r_t^* is the index return and η_t is the noise component of the index return in period *t*. Based on equation (7), the standard deviation of the index return, $\sigma_{r_t^*}$, which representing the volatility of the index (here named as *Vol*), and the 1st order autocorrelation coefficient, ρ_{r*} (here named as *AC(1)*), can be calculated as:

$$Vol = \sigma_{r_t^*} = \sqrt{\sigma_r^2 + \sigma_\eta^2} \qquad \textbf{(8)}$$

$$AC(1) = \rho_{r*} = (\rho_r \sigma_r^2 - \sigma_\eta^2 / 2) / (\sigma_r^2 + \sigma_\eta^2) \qquad \textbf{(9)}$$

Where σ_r^2 and σ_η^2 are the variance of the true return and the noise respectively, ρ_r is the 1st order autocorrelation coefficient of the true return.

Smaller σ_η^2 means less noise, a better estimation of market return. Thus, smaller *Vol* or larger *AC(1)* will indicate a better quality housing price index. We calculate these two statistics for each of the indices we have estimated in section 4.2. The results are shown in Table 5.

*** Insert Table 5 about here ***

From the table we can see that the ps-RS index has the lowest volatility and the highest first order autocorrelation among all the three indices, while the median price index has the highest volatility and lowest AC(1). These results suggest that the ps-RS has less noise, and the median price index has the most noise, among the three. This is also suggested perhaps more compellingly by a simple visual comparison of the three indices in Figure 2. The ps-RS index is noticeably smoother than the others, and the median index has bounces around the most. The superior performance of the ps-RS index in terms of low noise is probably due primarily to the much greater estimation sample size, created by the sales matching process that generates the pseudo-pairs.

5. Conclusion

The repeat sales model can be regarded as an extreme case of a matching rule, pairing only sales of the exact same house. We develop a pseudo repeat sales (ps-RS) model that is particularly appropriate in China's new residential market where each residential complex typically contains thousands of nearly homogeneous housing units sharing the same location and neighborhood attributes. We generate within-complex pairs. By regressing the price differential onto the classical RS time-dummies and the relatively small and easily observed within-pair differentials in physical attributes (the more problematical location and community variables are cancelled out), we are able to construct a ps-RS price index for new homes. This new index can effectively mitigate the problem of omitted variables which can bias hedonic index estimation. It also addresses the problems of the classical repeat sales index regarding sample size and sample selection bias. And it provides a parsimonious, simpler more transparent and easily understood specification for application in the real world in the Chinese context.

We estimate both the ps-RS index and a comparable hedonic index using a large-scale new home transaction dataset in Chengdu. The two indices show very similar trend and turning points, and intertwine with each other, suggesting that the hedonic index is not superior to the ps-RS index in terms of systematic results, while the ps-RS is simpler and more robust. Furthermore, the ps-RS index has a smaller volatility and larger first-order autocorrelation, providing a smoother, more noise-free index, based on the same underlying transaction sample. Thus, the ps-RS would seem to be an important new real estate price index methodology contribution particularly appropriate for rapidly urbanizing countries such as China. Actually the National Bureau of Statistics of China is now collecting micro housing transaction data (instead of relying on developers' self-reported numbers) and trying to develop a more reliable and also practical price index compiling methodology. Our methodology may be a good contribution to them.

Acknowledgement: Siqi Zheng is supported in part by the National Natural Science Foundation of China under Grant 70973065 and by Tsinghua University Initiative Scientific Research Program. Hongyu Chen and Yaoguo Wu provide excellent research assistance.

Reference

[1] Bokhari S, Geltner D. Estimating real estate price movements for high frequency tradable indexes in a scarce data environment[J]. The Journal of Real Estate Finance and Economics, 2010: 1-22.

[2] Case B, Pollakowski H O, Wachter S M. On choosing among house price index methodologies[J]. Real estate economics, 1991, 19(3): 286-307.

[3] Case B, Quigley J M. The dynamics of real estate prices[J]. The Review of Economics and Statistics, 1991: 50-58.

[4] Case K E, Shiller R J. The efficiency of the market for single-family homes[J]. 1989.

[5] Case K E, Shiller R J. Prices of single family homes since 1970: New indexes for four cities[J]. 1987.

[6] ClappJ, Giacotto C. Estimating price indices for residential property: A comparison of repeat sales and assessed value methods [J]. Journal of the American Statistical Association, 1992, 87: 300-306.

[7] Court A. Hedonic price indices with automotive examples [J]. The Dynamics of Automobile Demand, 1939, General Motors Corporation.

[8] Deng Y, Mcmillen D P, Sing T F. Private residential price indices in Singapore: A matching approach[J]. Regional Science and Urban Economics, 2011.

[9] Englund P, Quigley J M, Redfearn C L. The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition[J]. The journal of real estate finance and economics, 1999, 19(2): 91-112.

[10] Fu Y, Tse D K, Zhou N. Housing choice behavior of urban workers in China's transition to a housing market[J]. Journal of Urban Economics, 2000, 47(1): 61-87.

[11] Gatzlaff D H, Haurin D R. Sample selection and biases in local house value indices[J].

Journal of Urban Economics, 1998, 43(2): 199-222.

[12] Geltner D, Pollakowski H. On the Magnitude of Noise in the Moody's/REAL Index Return Reports, MIT Center for Real Estate-CREDL, 2008.

[13] Glendinning M, Muthesius S, Paul M C F S. Tower Block: Modern Public Housing in England, Scotland, Wales, and Northern Ireland[M]. Paul Mellon Centre for Studies in British Art, 1994.

[14] Griliches Z, Adelman I. On an index of quality change [J]. Journal of the American Statistical Association, 1961, 56:295, 535-548.

[15] Hansen J. Australian House Prices: A Comparison of Hedonic and Repeat‐Sales Measures*[J]. Economic Record, 2009, 85(269): 132-145.

[16] Kain J F, Quigley J M. Measuring the value of housing quality[J]. Journal of the American Statistical Association, 1970: 532-548.

[17] Mcmillen D P. Price indices across the distribution of sales prices: A matching approach[J]. Urbana, 2010, 51: 61801.

[18] Quigley J M. A simple hybrid model for estimating real estate price indexes[J]. Journal of Housing Economics, 1995, 4(1): 1-12.

[19] Rosen S. Hedonic rrices and implicit markets: product differentiation in pure competition [J]. Journal of Political Economy, 1974, 82:1, 34-55.

[20] Wallace N E, Meese R A. The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches[J]. The Journal of Real Estate Finance and Economics, 1997, 14(1): 51-73.

[21] Zheng S, Fu Y, Liu H. Housing-choice hindrances and urban spatial structure: Evidence from matched location and location-preference data in Chinese cities[J]. Journal of Urban Economics, 2006, 60(3): 535-557.

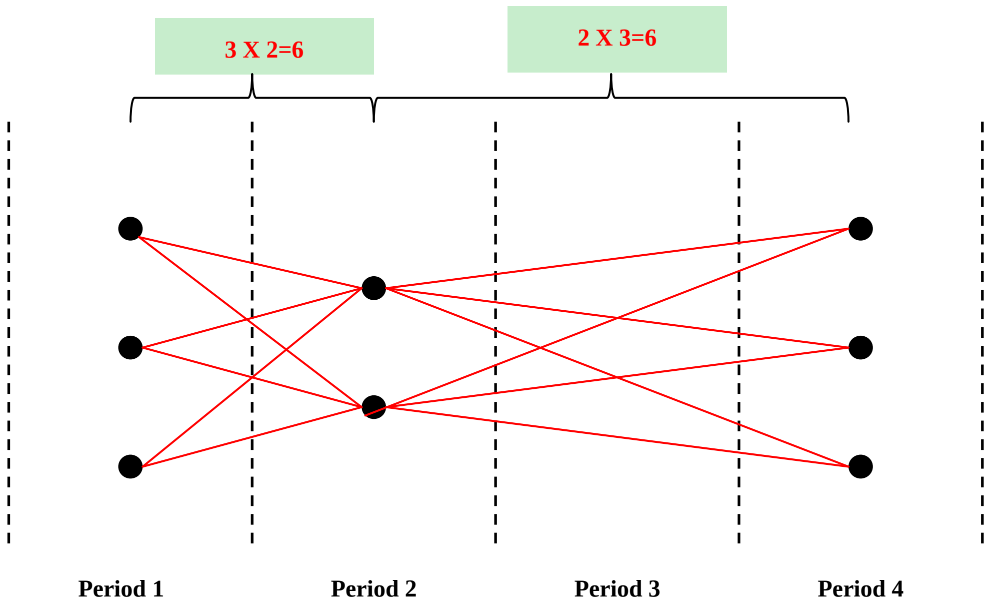

Figure 1 Matching Process across Periods within a Complex

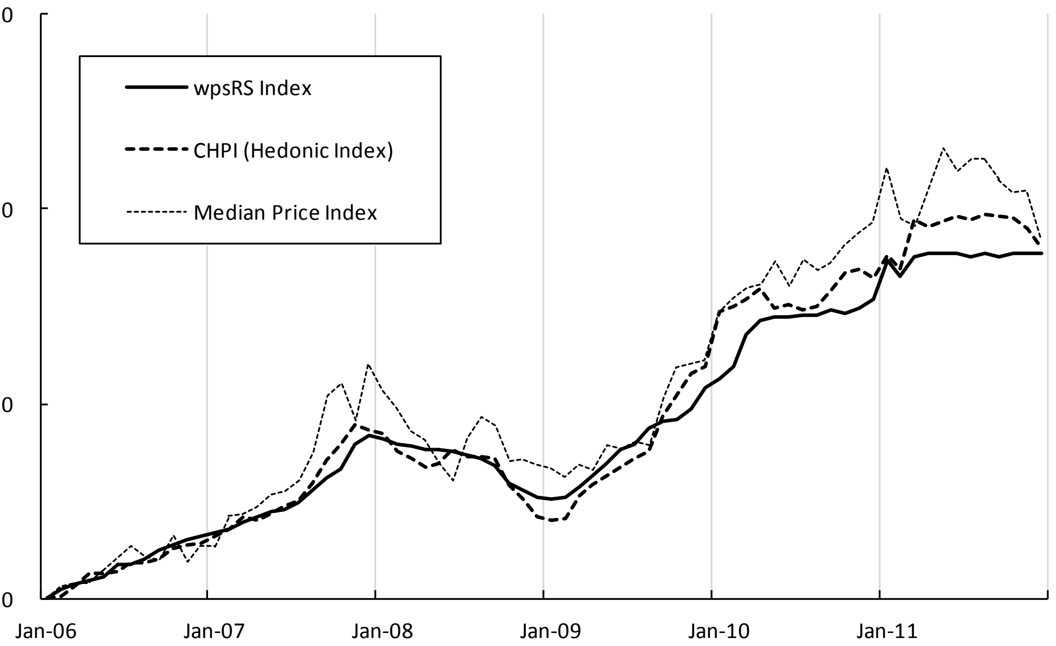

Figure 2 Comparison of different housing price index

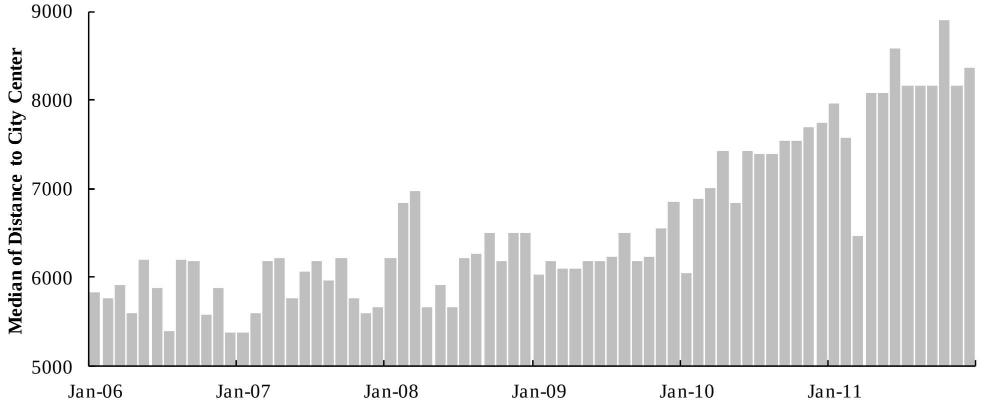

Figure 3 Suburbanization of Residential Development

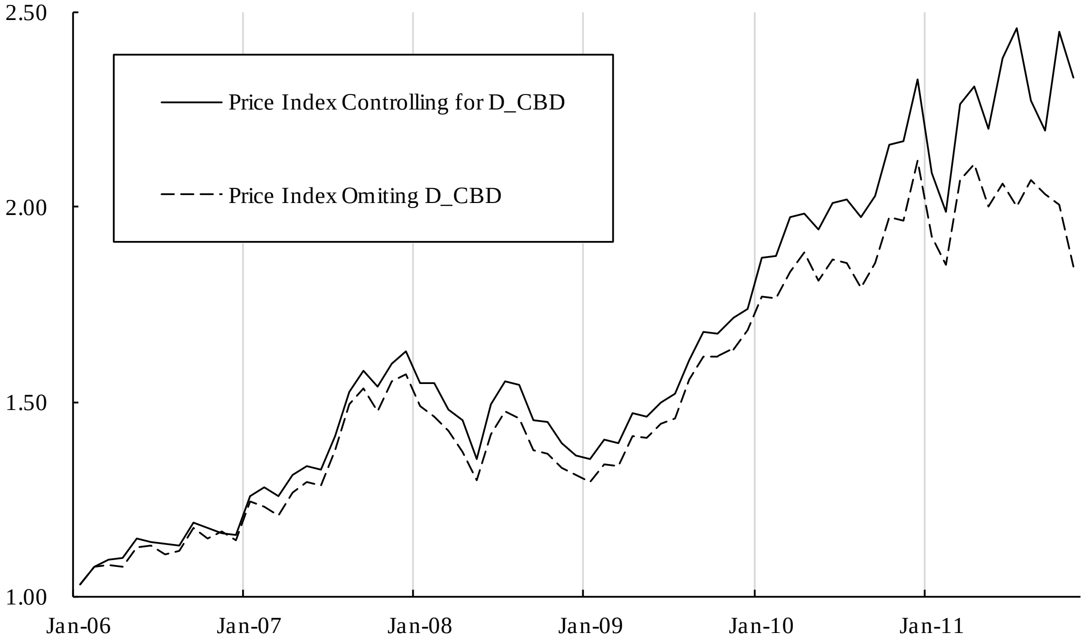

Figure 4 Comparison of Two Hedonic Indexes

(with and without D_CBD as a control variable)

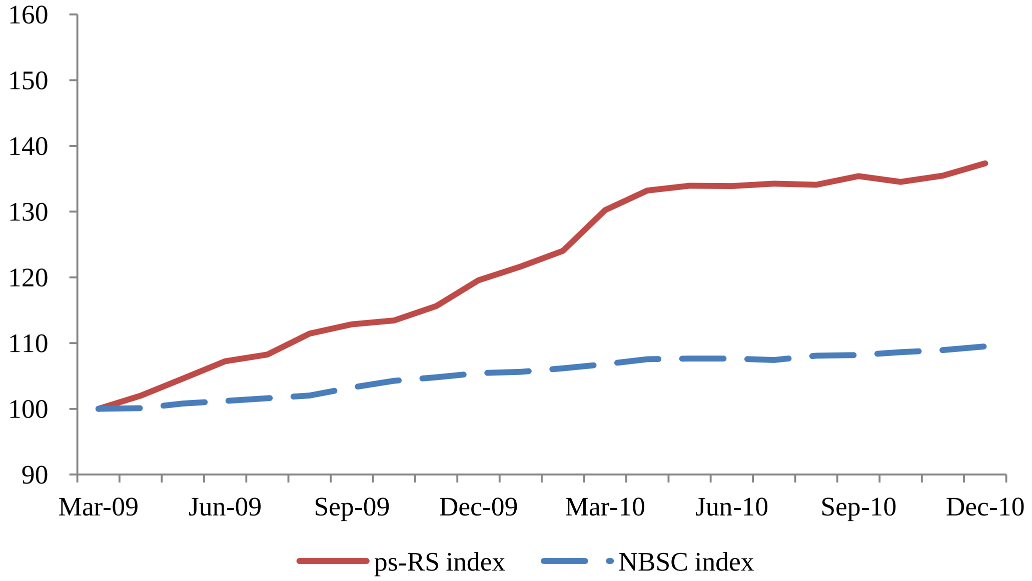

Figure 5 Comparison of ps-RS index and NBSC index

Table 1 Variable Definition and Descriptive Statistics

Variables	Unit	Description	Mean	Median	Max	Min	Sd.Dev
Physical Attributes							
PRICE	million RMB Yuan	Total purchase price	0.58	0.51	18.50	0.05	0.34
SIZE	square meter	housing unit size	98.61	89.35	735.37	14.79	33.24
FLOOR	/	Floor number	12.41	11.00	55.00	1.00	7.98
BEDROOM	/	number of bed rooms	2.24	2.00	9.00	0.00	0.79
BATHROOM	/	number of bath rooms	1.39	1.00	9.00	1.00	0.54
TFLOOR	/	total floors above ground	23.67	22.00	66.00	3.00	8.22
Location Attributes							
D_CBD	km	distance to city center	6.96	6.50	36.01	0.26	3.09
ZONE	dummy	33 zones					

Table 2 Descriptive Statistics of Matching Results

Variable	Mean	Median	Max	Min	Std. Dev
Average number of transactions in a matching space (complex) before matching	214	163	2,191	1	218
Average number of pairs generated in a matching space (complex) after matching	10,851	3,398	591,419	1	27,371

Table 3 Estimate results of Ps-RS model

Variables	Coefficient (t-statistic)
$\Delta\ln(SIZE)$	1.02 (12000***)
$\Delta\ln(FLOOR)$	0.01 (615.96***)
$\Delta BEDROOM$	0.001 (21.39***)
$\Delta BATHROOM$	0.02 (348.22***)
Month Dummies	Yes
R^2	0.874
Obs.	22,281,758

t statistics in parentheses

$^{*}p< 0.10$, $^{**}p< 0.05$, $^{***}p< 0.01$

Standard errors clustered by complex.

Table 4 Estimate Result of Hedonic Model

Variables	Coefficient (t-statistic)
ln(*SIZE*)	1.16 (456.15***)
ln(*FLOOR*)	0.01 (28.42***)
BEDROOM	-0.09 (-104.49***)
BATHROOM	0.08 (72.57***)
ZONE Dummies	Yes
Month Dummies	Yes
Intercept	7.84 (727.27***)
R^2	0.708
Obs.	469,070

t statistics in parentheses

*p< 0.10, **p< 0.05, ***p< 0.01

Standard errors clustered by complex.

Table 5 Quality Judgment of Indices

Indices	*Vol*	*AC(1)*
Ps-RS Index	0.00055	0.00014
Hedonic Index	0.00129	0.00004
Median Price Index	0.00361	-0.00030