A new approach for constructing home price indices: The pseudo repeat sales model and its application in China

Citation: Guo, Xiaoyang et al. "A New Approach for Constructing Home Price Indices: The Pseudo Repeat Sales Model and Its Application in China." Journal of Housing Economics 25 (September 2014): 20–38.

As Published: http://dx.doi.org/10.1016/j.jhe.2014.01.005

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/111137

Version: Original manuscript: author's manuscript prior to formal peer review

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)

* Corresponding author: Siqi Zheng
Email address: zhengsiqi@tsinghua.edu.cn
Mail address: Department of Construction Management, He Shanheng Building, Tsinghua University, Beijing 100084, P. R. China
Phone number: +86-18611045757

Highlights

- We develop a "pseudo repeat sale price index" for newly-constructed homes in China.
- This index has advantages over the hedonic and the repeat-sales in China's case.
- Our empirical work using Chengdu data proves this methodology's strengths.
- This methodology can be applied in other rapidly developing countries.

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)

* Corresponding author: Siqi Zheng
Email address: zhengsiqi@tsinghua.edu.cn
Mail address: Department of Construction Management, He Shanheng Building, Tsinghua University, Beijing 100084, P. R. China
Phone number: +86-18611045757

This version: January 2013

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 in Chinese cities (87% in 2010). As a result, the standard repeat sales approach cannot be employed, as a new housing units 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 rapid building quality improvement.

Taking advantage of the uniquely large scale and homogeneous nature 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 often contains several phases and 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 so forth. Based on specific criteria, we match two very similar new sales within three versions of a defined matching space: within a complex, within a phase of a complex, or within an individual building, respectively. We thus 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 across time between the two sales in each pseudo-pair 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, thereby cancelling out or controlling for locational and community variations, we are able to construct a ps-RS price index for new homes. We examine three versions of such an index: complex-based, phase-based and building-based.

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 and sample selection bias, as we are effectively able to use all sales. Another advantage of this index is its transparency and ease of understandability in the Chinese context, thus allowing 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(444,596 observations) in Chengdu, Sichuan Province. We estimate our ps-RS indices and compare them with a corresponding standard hedonic index. The two types of indices show very similar trend and turning points. The complex-based version of the ps-RS index essentially parallels the hedonic index, suggesting that the hedonic index is not superior to that version of the ps-RS index in terms of systematic results. The phase-based version of the ps-RS index has a lower growth trend and the building-based version lower still. This indicates that the hedonic index and the complex-based ps-RS index do not sufficiently control for omitted variables relating to the physical quality of the units, which in China has been improving very rapidly, and it suggests that the building-based version of the ps-RS index provides the greatest control for such quality differences. Compared to the hedonic, all of the ps-RS indices have less volatility, greater first-order autocorrelation, and smaller deviation from a Hodric-Prescott smoothed benchmark index, suggesting that the ps-RS models exhibit less random estimation error (or "noise").

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

JEL Code: R3, R31

A New Approach for Constructing Home Price Indices in China: The Pseudo Repeat Sales Model

1. Introduction

In the world of transaction price indices used to track the dynamics in housing markets, 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 location, 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, the FHFA and S&P/Case-Shiller home price indices).

Consider two unique features in China's urban residential market. First, 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. Thus, the classical repeat sales approach is of very limited usefulness because the typical housing unit in China has only appeared 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 very rapid evolution of urban spatial structure, infrastructure construction, and (most difficult to observe) the quality and features and amenities within the housing units themselves (such as apartment design, appliances, finishes, and HVAC) as household income rises at an extremely rapid rate. Secondly, housing development in China occurs at a uniquely large scale in terms of numbers of units developed at once, and with correspondingly widespread homogeneity in the units.

The proposal in this paper is to develop a new type of "repeat sales" model, which we dub "pseudo repeat sales" (ps-RS). Essentially, we propose a new matching criterion that is particularly appropriate in Chinese cities. We deal with the omitted variables issue by employing a within-building matching criterion instead of the more stringent, classical same-unit criterion[1]. 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 paired units are not identical) of the type that we 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 relies only on variables for which good data can be easily obtained. We believe the ps-RS still retains essentially the characteristics of a "repeat sales" model. In this paper we present an argument and evidence that the ps-RS can produce a more reliable and practical housing price index which is especially suitable for the new residential markets in Chinese cities.

The rest of this paper is organized as follows: Section Two will present some relevant background and literature review. Section three 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 describe in detail our approach for developing the ps-RS index in Section Four. After data description in Section Five, the index calculation results for our demonstration city of Chengdu are presented in Section Six, including 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 for new housing). Section Seven concludes.

[1] The matching criterion can also be applied to sales within the same complex, or the same sale phase. However, as we will discuss below, larger matching spaces appear to be less effective in mitigating the problem of omitted variables and controlling for quality differences. Our empirical results indicate that the within-building criterion is best in Chengdu.

2. Background & Literature Review

The hedonic approach goes back to Kain and Quigley (1970), who decomposed the components of housing price dynamics using the hedonic model, from which a housing price index was 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 alternative methods have been proposed to construct a 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 those 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, 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, as demonstrated by Clapp and Giacotto (1992). 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.[2]

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

[2] 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).

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).[3]

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. 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 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, which is relevant to the current paper, is to view the repeat sales specification as one (extreme) solution to a matching problem. The objective is to match or pair sale observations together according to certain specific criteria so as to cancel out unobservable attributes, making the model more parsimonious and robust so that it does not need as much 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 renovations in the

[3] 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.

neighborhood or improvements in the house). McMillen (2010) suggests a more open matching approach as an alternative. Deng, McMillen and Sing (2011) expanded this approach and applied it to Singapore's residential market. They predict the sale probabilities of all the transactions, and then match each transaction after the base period with a transaction in the base period that has 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 relatively 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 has much better data and lacks some of the extreme modeling challenges found in Chinese cities. Wu et. al. (2012) compare the performances of the simple average method, the matching approach, and the hedonic modeling approach in estimating housing price indices in a Chinese city. Their results show that the hedonic approach works the best during their study period in that city.

3. 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 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). On the other hand, 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. The omitted (positive) quality variables will lead to an upward biased index.

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'

[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). According to Real Capital Analytics, land sales transactions (ground leases) of over USD 10 million totaled over USD 250 billion in China in 2011. The comparable figure in the U.S. in the same year was less than $10 billion (down from over $30 billion in 2007).

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 couple of high-rise condominium buildings that share nearly the same location attributes, common architectural design, structure type and community/property services. A large complex may be divided to several phases, and those phases are developed and sold sequentially. Each phase contains a couple of multi-storied or high-rise buildings. A small complex usually has one phase and all buildings are built at the same time. There are small within-complex differences across phases or buildings such as the sale start time, whether facing the main street (noise), distance to the complex's main entrance, etc. The within-phase differences are even smaller. The housing units within a single building are the most homogenous except of the small differences in floor number (height above the ground within the building), unit size, number of bedrooms, and the direction the main bedroom faces. Relatively reliable data exists for these attributes. These circumstances therefore provide a unique opportunity to develop a "pseudo repeat sales" (ps-RS) model.

In the ps-RS method we match two very similar new sales within a building (or within a phase, or within a complex, depending on the definition of the matching space). We thereby create a paired sale observation. We call these pairs "pseudo repeat sales" (or "pseudo pairs") 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 within-pair price differential between the first and second sales 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 attributes variables, are cancelled out of the model just as they are 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 in China.

4. Index Construction Methodology

In this section we describe the ps-RS methodology in detail. After describing the matching process to construct the pseudo-pairs, we present the regression specification and then we address a data weighting issue that arises with the methodology.

4.1 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 narrow matching space implicitly restricts the matching rule to be the same location and physical attributes (except for age and possible renovation).[6]

[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. new home price evolution.

[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).

In our pseudo repeat sale model, we expand the matching space from one house to three possible alternative larger definitions (from larger to smaller): a residential complex, a phase within a complex, or a building within a phase. For each version of the matching space, we construct a ps-RS index. As mentioned above, all housing units in a complex share very nearly the same location and neighborhood attributes, and a subset of physical attributes. If a complex contains several phases, each phase will have a specific "market entrance" date on which day all units in that phase become available on market. A possibility is that units in the first phase may be sold at a price discount because the buyers face higher uncertainty and have to bear noise and dust pollution when other later phases are under construction, and the developer may be particularly eager at that point to establish the viability of the project. Wu et. al. (2012) also discuss the developer's pricing strategy when setting the prices for units in different phases within a complex. In fact, a hedonic regression shows that the first phase does have a price discount of about 4.8%, but there is no significant discount for later phases. To mitigate this first-phase effect, we drop all the transactions in the first phase in all complexes when we construct the complex-version of the ps-RS index.

Any two units in a within-phase pair share the same "market entrance" date, so we don't need to worry about a first-phase effect for the within-phase ps-RS. The units in the within-phase pseudo-pairs also exhibit more commonality in a larger subset of attributes than those in the within-complex pseudo-pairs. And of course the units in the within-building pseudo-pairs have even more commonality.

Applying within-pair first differencing will cancel out any variables for which the attributes are the same between the two units, including both observable and unobservable attributes. Only attributes that differ between the two units within a pair will be left on the right-hand side as independent variables, differenced between the second minus the first sale, reflecting the "hybrid" specification of repeat sales and

hedonic modeling. *A* priori we prefer the building-version of the ps-RS index because it can to the highest degree mitigate the omitted variables problem. However, in reality, if the index compiling authority does not have the phase identifier (or the building identifier), the best it can do is to construct the complex-version (or phase-version) of the ps-RS index. Since we have both phase and building identifiers for the Chengdu database we use in this paper, we will construct all three versions of the ps-RS indices, and do some comparisons among them.

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 pair construction rule that we use is to match one transaction with its most temporally adjacent transaction in the same matching space. Suppose we have four periods in total. Taking the within-building version of the ps-RS index as an example, suppose that in a given building there are 3 transactions in the 1st period, 2 transactions in the 2nd period, 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 there is no transaction in the 3rd period, when we 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 from the 8 sales that have occurred.

*** Insert Figure 1 about here ***

We do not match the transactions in the 1st period directly with those in the 4th period into pairs because they are not "adjacent" transactions. The rationale behind is that including "non-adjacent" transaction pairs would be redundant from an information perspective and generate an excessive quantity of data. The price change between the 1st and 4th periods is fully reflected in the price change between the 1st and the 2nd

periods plus that between the 2nd and the 4th periods .

Though the subject building in our example has no transaction in the 3rd period, another building may have some transactions in that period. Since the whole index sample consists of thousands of complexes, every period will be amply included in the index estimation sample.

4.2 Regression Model

The standard hedonic model to construct a housing price index is shown as Equation (1) (Quigley, 1991), where P_i is 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, $D_{t,i}$ is the time dummy which equals 1 if the sale occurs in period t, 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. We again use the building-version as the demonstration. Here buildings are indexed by j, periods (months) are indexed by t. Within building 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.

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

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].

4.3 Weighting Adjustment

In Equation (2) the observation is a pseudo-pair. A potential problem is that in the generation of the pseudo-pair estimation sample, the original sample size distributions over time and across buildings (or complexes/phases) will be changed, relatively speaking compared to a corresponding hedonic index. Consider two adjacent periods *r* and *s*, and suppose there are N_r and N_s observations in these two periods in a representative building, 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 ps-RS 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 relative to the hedonic. This is also true across phases or complexes.

We therefore introduce a weighted OLS procedure 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 building *j*, the weight is:

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

An alternative weighting procedure would be equal weighting – setting the weight formula so that each time period has the same weight. After all the pairs are generated, the weight applying to the pairs in which the latter transaction occurs in period *s* is:

[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. (And the time dummies in a classical same-house RS model also reflect the aging of the house, something that the ps-RS does not reflect as all the houses are new).

$$W_s = 1/\sum_{j=1}^{Q_s} N_{r_j,j} \cdot N_{s,j} \qquad (r_j<s) \qquad (4)$$

Where $N_{s,j}$ is the number of transactions in period s in building j, and $N_{r_j,j}$ is the number of transactions in period r_j in building j. Period r_j is the most adjacent previous period to period s (in different buildings, this "most adjacent previous period" may be different). Q_s is the total number of buildings in period s.

There is no general rule for which weighting adjustment is the best one. In principle the first rule should be most appropriate for comparing the ps-RS index with a corresponding hedonic index, and that is the result we will report in this paper. However, in fact we have examined ps-RS indices under both of the above two weighting schemes. The results are nearly identical. However, the second weighting scheme (equal-weighted periods) produces an index that tracks very slightly below the first weighting scheme.

5. Index Estimation and Discussion

We test the ps-RS index method on a dataset of new residential unit transactions in Chengdu, the capital city of Sichuan Province. The Chengdu local authority provided us a high quality micro data set of all transactions in its new housing market, making it possible to estimate a relatively good hedonic index. It thus presents a good laboratory to explore the ps-RS method because we can compare it to a relatively good hedonic index. In this section we describe the data as well as our estimation results including a comparison with a classical hedonic index.

5.1 Data

The Chengdu dataset is very large (and in this respect is not untypical of what Chinese cities can provide). The database contains the full records of Chengdu's new

residential sales from January 2006 through December 2011, consisting of 2152 complexes and altogether 444,596 housing units after data cleaning.[8] 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 among the 33 zones[9] defined by the Chengdu Local Housing Authority). Table 1 shows the descriptive statistics of these variables.

*** Insert Table 1 about here ***

5.*2 Index Estimation Using ps-RS Model*

We have three versions of matching space for our ps-RS model: complex, phase and building. The larger the matching space is, the more pseudo-pairs can be generated. For the complex-version, 31.6 million pairs are generated from the 444.6 thousand transactions in 901 complexes.[10] For the phase-version, 22.3 million pairs are generated in 2,174 phases. For the building-version, 14.4 million pairs are generated in 3,913 buildings.

Equation (2) is regressed over all the pseudo-pairs using WLS, with standard errors clustered by the corresponding matching space. Table 2 reports the estimated results

[8] We drop those "outlier" observations with extreme price per square meter (the 0.1% highest and the 0.1% lowest). We also drop those transactions whose time on market (TOM) exceeds the 95 percentile in its distribution at the phase level. In effect, we're assuming a "natural vacancy rate" of 5%. 24,474 observations are dropped, which is about 5.21% of the original sample size (469,070 observations).

[9] We divide the urban space of Chengdu into 33 zones by two rules: the ring-road and the direction. Chengdu is a monocentric city, with four main ring-roads including the inner ring-road in the central city and another three ring-roads successively from inside to outside named as the 1st, the 2nd and the 3rd ring 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.

[10] To control for the first-phase effect, we drop the transactions in the first phase when we estimate the complex-based ps-RS regression. There is no first-phase effect for the phase-based or building-based ps-RS regressions.

of the building-version, phase-version and complex-version ps-RS models, respectively. As explained above, on an *a priori* basis we prefer the building-version regression because it can mitigate the omitted variables problem to the highest extent. All the coefficients of the physical attributes in the three regressions are statistically significant and have the expected signs. The ps-RS model can explain 90.28%, 85.68% and 81.32% of cross-pair differences in price growth in the building-version, phase-version and complex-version ps-RS regressions. Based on the coefficients of the time dummies, the three versions of ps-RS Indices are calculated and shown in Figure 2. We also estimate the standard hedonic price index based on the same sales transactions dataset (with zone dummies to control for location attributes, see Table 3 for regression results), and we show it also in Figure 2 for comparison. Since we want to compare our ps-RS indices with the hedonic index, we employ the first weighting scheme described in Section 4.

*** Insert Table 2 about here ***

*** Insert Table 3 about here ***

*** Insert Figure 2 about here ***

The black line with dots is a hedonic price index calculated based on the hedonic regression shown in Table 3. The red solid line, red short-dashed line and red long-dashed line are the complex-version, phase-version and building-version ps-RS indices, respectively. We can see that the ps-RS indices and the hedonic index have a similar overall trend and similar turning points. Before mid-2007, all 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.

As stated above, there are two broad categories of omitted variables – location attributes and physical attributes. On one hand, the rapid urbanization in Chinese cities 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. This will cause a downward bias in a hedonic index.

On the other hand, with such rapidly rising per capita income in Chinese cities, 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. 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 size and 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. The above logic is also true when we compare different versions of ps-RS indices. More physical quality variables (observed and unobserved) can be cancelled out and effectively controlled for when we estimate the ps-RS index with smaller matching space.

In Chengdu's case, the complex-based version of the ps-RS index intertwines with the hedonic index, essentially paralleling it. This implies that for the Chengdu dataset the potential problem of omitted location variables is in fact not a serious problem in

practice. Since the within-complex ps-RS index does control quite well for omitted location variables, and the ps-RS index essentially tracks the hedonic index, apparently the 33 zones in the hedonic index are controlling quite well for location effects in the pricing.[11] However, the ps-RS index is much smoother (with smaller volatility) than the hedonic index. This suggests that the ps-RS index is better (as will be discussed further below).

Unlike the complex-based ps_RS, the phase- and building-based versions of the ps-RS indices do reveal a systematic difference from the hedonic index. Thus, the type of physical quality attributes that the hedonic index and the complex-based ps-RS index cannot control for as well as the phase- and building-based indices do apparently affect the pricing trend. In particular, the phase- and building-based indices both tend to track below the complex-based and hedonic indices. As between the phase- and building-based indices, before 2009 the two track together. But after 2009 the phase-based index increases faster than the building-based. Since the phase-based index cannot do as good a job of controlling for omitted physical quality variables as the building-based index, it appears that improvement over time in omitted physical quality variables impart a positive bias into the phase-based index, at least in the case of our Chengdu dataset.

Apart from dealing with omitted location and physical quality attributes, there are two other sources of difference between the ps-RS and hedonic indices, which may partly explain the differences we observe in Figure 2 between the hedonic versus ps-RS indices. 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. For example, the building-based ps-RS index is estimated on 14.4 million observations, while the hedonic is estimated on less than a half million. This larger sample size should help the ps-RS model to be estimated more precisely, resulting in less noise in the index,

[11] Of course, this might not be the case in all cities.

giving the index a smoother appearance. Another 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 could in theory affect the results. As noted in Section 1, an additional practical advantage of the ps-RS model over the hedonic approach may be greater ease of understandability or communication to practitioners and policy makers.

To provide more background information, here we also compare our building-based 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 3 shows the two indices for Chengdu from 2009M3 to 2010M12 (we are only able to find systematic NBSC index series for this period). The NBSC index was calculated by simply averaging developers' self-reported price changes compared to the 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. Wu et. al. (2012) discuss the shortcomings of this NBSC index in detail. We can see that in Figure 3 the NSBC index tracks significantly lower than our ps-RS index (of course also much lower than the hedonic index).

*** Insert Figure 3 about here ***

5.3 Judging Index Quality

There are two broad categories of errors of most potential concern in housing price indices – systematic bias and random error. As we discussed above in Section 5.2, the building-based ps-RS index does a better job in mitigating the omitted variables

problem which is the major likely cause for systematic bias in a transaction price based index such as the present context.[12] But what about random estimation error? This section reports two formal tests of the quality of the ps-RS indices in terms of their reliability, an out-of-sample prediction test and a smoothness test against random noise.

5.3.1 Out-of-sample robustness check

We randomly divide the whole sample into two sub-samples with the same sample size.[13] There are no overlapping data points between the two 50% random sub-samples. We estimate two separate ps-RS indices for the two sub-samples. The two indices are almost the same (no visually apparent difference at all, as seen in Figure 4). The correlation between the two indices is 0.999993. Furthermore, we conduct the mean-comparison test between those two indices with the null hypothesis as $mean(Index_{1,t} - Index_{2,t}) = 0$, where t indicates the period number. The t-value of the test is only 0.015 and the p-value is as high as 0.988, which indicates that statistically we cannot reject the H0 so that the two indices based on randomly constructed two 50% sub-samples are almost the same with each other.

*** Insert Figure 4 about here ***

5.3.2 Comparing indices regarding random error

In this section we explore three different tools to compare the three versions of ps-RS

[12] Of course, bias can be caused by sample selectivity or unbalanced data sourcing. However, in the present context the dataset consists of virtually all new residential sales in Chengdu. This is not to say, however, that an aggregate index such as we are here examining would necessarily be a good representation for all submarket segments. But the estimation sample size is large enough to allow considerable construction of sub-indices to examine sub-markets. In lower frequency transaction-based indices smoothing and lagging bias can be caused by temporal aggregation in the time-dummy variables unless explicitly corrected. However at the monthly frequency we're employing this would not seem to be a significant concern as there is relatively little real estate price movement within each month.

[13] We assign a uniform-distributed random number between 0 and 1 to each observation using the command "runiform" in STATA. We then assign the observations with this random number less than 0.5 to the first sub-sample and the others to the second sub-sample.

indices, as well as the hedonic index, in terms of random statistical estimation error, the type of error that can impart "noise" into the index.[14] Geltner and Pollakowski (2008, as reported in Bokhari and Geltner, 2010) describe a model of index noise which suggests two indicators that will often 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

[14] 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.

accumulate over time. For an index beginning *T* periods ago, we have:

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

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 Figure 2. The results are shown in the first two rows in Table 4. We can see that the volatility measures of the three ps-RS indices are much lower than that of the hedonic index. The ps-RS indices also have much higher first order autocorrelation coefficients than the hedonic index. Among the three ps-RS indices, the building-based version has the lowest volatility and the highest first order autocorrelation. These results suggest that the ps-RS has less noise than the hedonic, and the smaller the matching space is, the better performance in terms of noise reduction. Presumably this is due to the better controlling for the variation in housing's characteristics, given a sample size that is already more than sufficient to mitigate random estimation error. This conclusion is also suggested perhaps more compellingly by a simple visual comparison of the

indices in Figure 2. The ps-RS indices are noticeably smoother than the hedonic. 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.

The third tool we use is a test based on the Hodrick & Prescott filter (HP filter). In some sense this is a formal quantification of the "eyeball test" of the visual smoothness of the index in a graph. The HP filter has been promoted by Hodrick and Prescott (1997) in order to analyze time series data. It is a spline fitting method that divides a time series into two smooth components, a secular trend and a cyclical component. The HP filter has been popularly used to analyze macroeconomic variables as well as the price series in the real estate market (Cocconcelli and Medda, 2013). In STATA 12.0, we separate the trend and cyclic series for all of our target indices in Figure 5: building-version, license-version and complex-version ps-RS indices, and the hedonic index. The smoothed trend and cyclical series of the four indices are shown in Figure 5 (red lines). The HP-based comparison of the indices is essentially a quantification of smoothness. We want to see which index has the least deviation from its smoothed HP representation. We compute the index returns and the smoothed returns of the corresponding HP representation. For each type of index we compute the sum of squared differences between the index returns and its smoothed returns across the history, and then we compare these sums. The results are shown in the bottom row in Table 4. Once again, the building-based version of the ps-RS index comes out looking best, with the smallest deviation from its smoothed representation. This is also consistent with our findings from the AC(1) and volatility tests which are also reported in the table.

*** Insert Table 4 about here ***

*** Insert Figure 5 about here ***

6. 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 several phases, a couple of buildings, and thousands of nearly homogeneous housing units sharing the same location and neighborhood attributes. We generate within-complex, within-phase and within-building pairs, respectively. 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 three versions of ps-RS price indices for new homes, with complex, phase and building as matching spaces, respectively. These new ps-RS indices show good results in mitigating the problem of omitted variables which can bias hedonic index estimation. The building-version ps-RS index does the best job in this regard because its within-pair differential is the smallest. Our ps-RS index also addresses the problems of the classical repeat sales index regarding sample size and sample selection bias, as it uses all available sales transactions. By actually increasing the effective estimation sample size through the sale-pairing process, the ps-RS results in very smooth, reliable indices. 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 indices and a comparable hedonic index using a large-scale new home transaction dataset in Chengdu. The two types of indices show very similar trend and turning points. The complex-based version of the ps-RS index parallels the hedonic index, 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 and provides a smoother index. The phase-based version of the ps-RS index has a lower growth trend, and the building-based version shows the lowest price growth of all. From these results we infer that omitted or poorly measured location

attributes do not have much net effect on a hedonic index in Chengdu, but omitted or poorly measured physical quality attributes tend to cause an upward bias in the price growth trend of the hedonic index (and the complex-based ps-RS index). Furthermore, the ps-RS indices have better performance than the hedonic index in tests for random error or estimation reliability, as indicated by out-of-sample prediction and tests of smoothness. 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. Recently the National Bureau of Statistics of China has been 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 construction methodology. The ps-RS method may warrant serious consideration.

Acknowledgement: Siqi Zheng is supported in part by the National Natural Science Foundation of China under Grant 70973065 and, and by Tsinghua University Initiative Scientific Research Program. Hongyu Chen and Yaoguo Wu provide excellent research assistance. The authors appreciate the suggestions of Albert Saiz as well as participants in the National University of Singapore Department of Real Estate Seminar in August, 2012.

Reference

[1] Bokhari S, Geltner D. Estimating real estate price movements for high frequency tradable indexes in a scarce data environment. 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. Real estate economics, 1991, 19(3): 286-307.

[3] Case B, Quigley J M. The dynamics of real estate prices. 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. 1989. NBER Working Paper No.2506.

[5] Case K E, Shiller R J. Prices of single family homes since 1970: New indexes for four cities. 1987. NBER Working Paper No.2393.

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

[7] Cocconcelli L, Medda F. Boom and bust in the Estonian real estate market and the role of land tax as a buffer . Land Use Policy, 2013: 30, 392-400.

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

[9] Deng Y, Mcmillen D P, Sing T F. Private residential price indices in Singapore: A matching approach. Regional Science and Urban Economics, 2012, 42(3): 485-494.

[10] 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.

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

[12] 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.

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

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

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

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

[17] Hodrick R, Prescott E. Postwar U.S. Business Cycles: An Empirical Investigation[J]. Journal of Money, Credit and Banking, 1997, 29:1, 1-16.

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

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

[20] Peláez R. The housing bubble in real-time: the end of innocence. Journal of Economics Finance, 2012: 36, 211-225.

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

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

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

[24] Wu J, Deng Y, Liu H. House Price Index Construction in the Nascent Housing Market: The Case of China. Working paper. Tsinghua University and National University of Singapore.

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

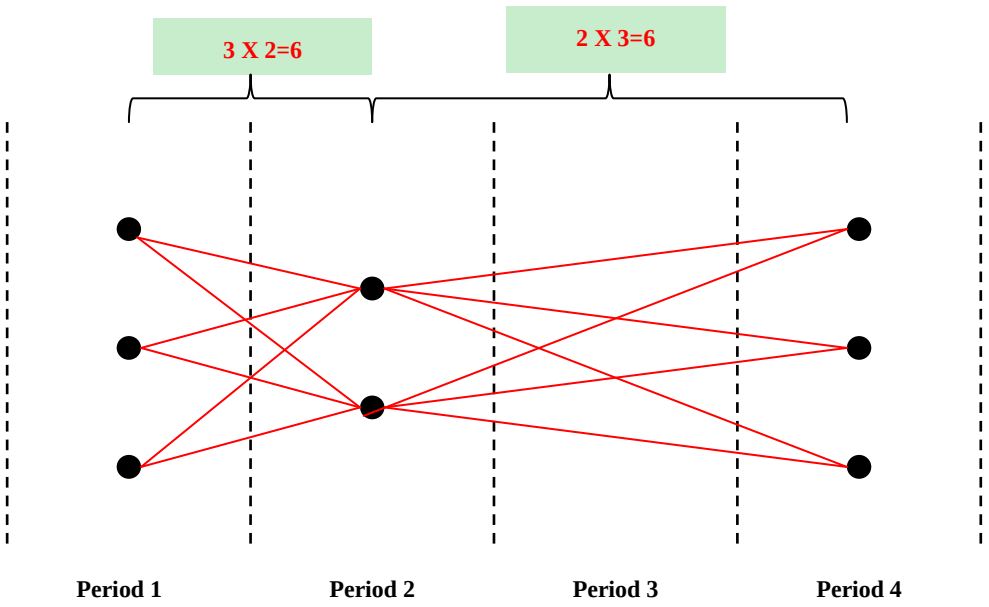


Figure 1 Matching Process across Periods within a Matching Space (Building, Phase, or Complex)

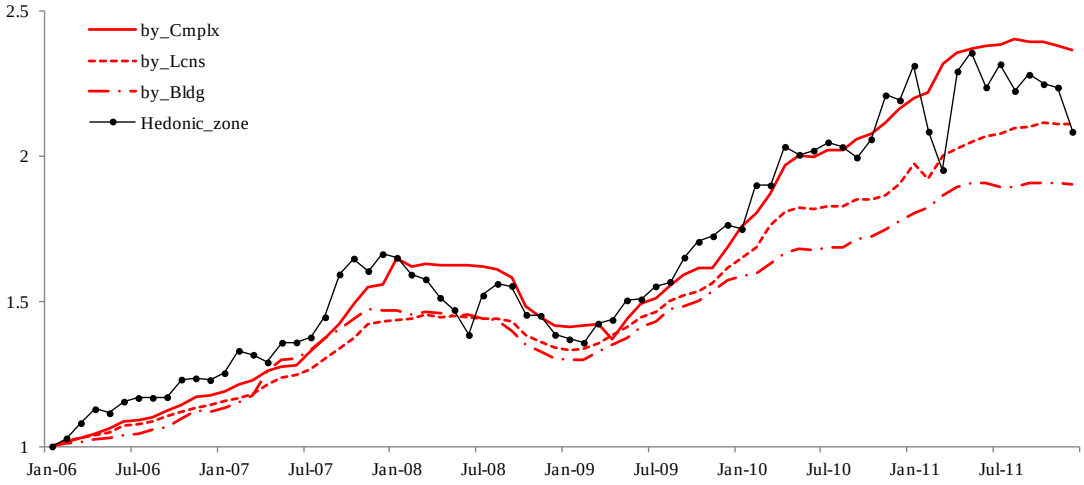


Figure 2 Three ps-RS Indices and the Hedonic Index for Chengdu

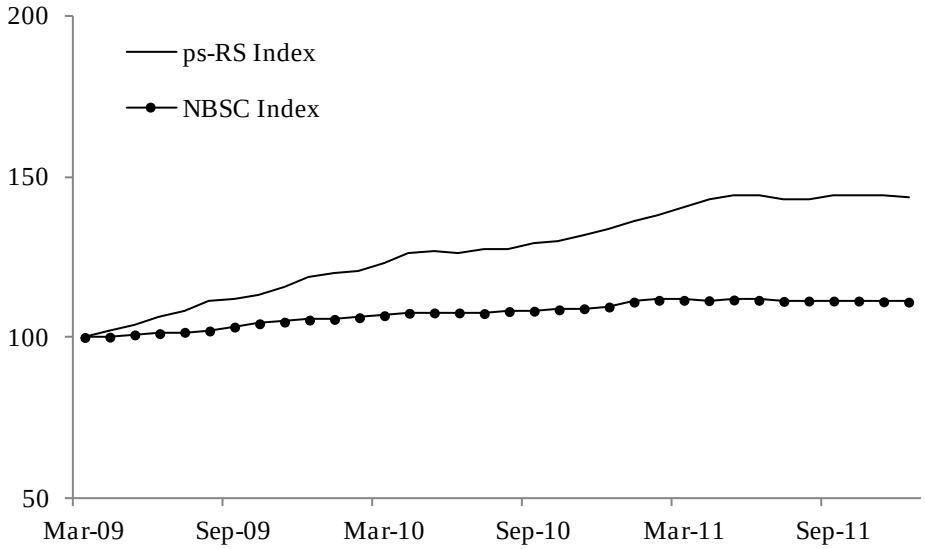


Figure 3 Comparison of ps-RS index and NBSC index

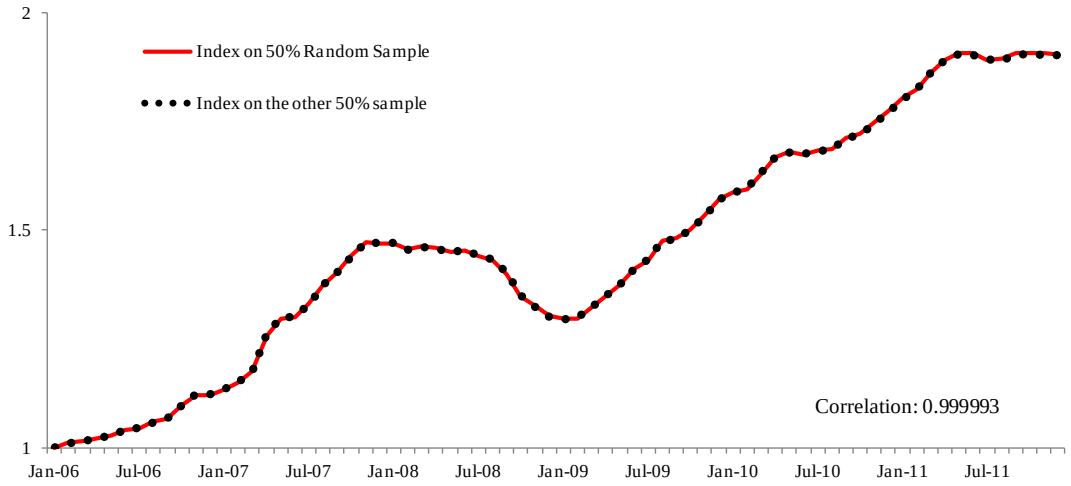
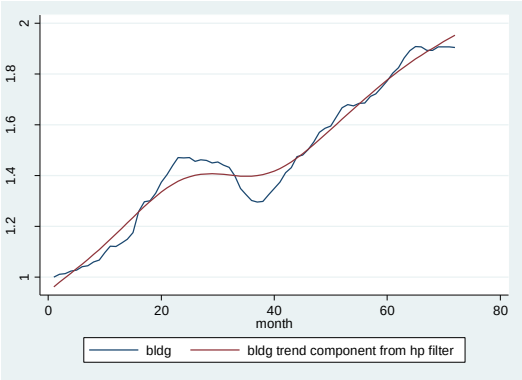
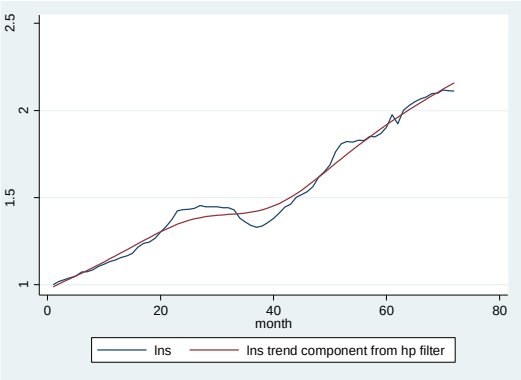


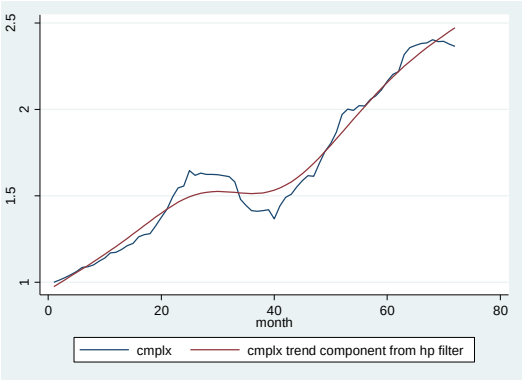
Figure 4 Out-of-Sample Robustness Check



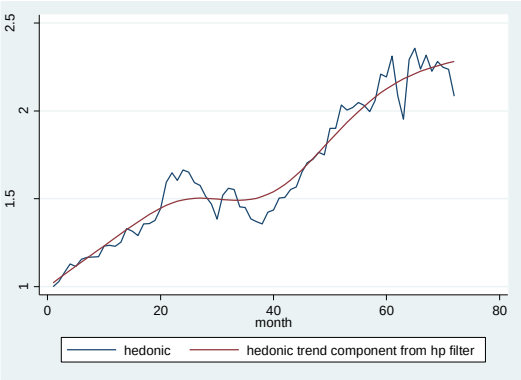
Building-version ps-RS index



Phase-version ps-RS index



Complex-version ps-RS index



Hedonic index

Figure 5 Trend series in four indices using HP filter method

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	0.57	0.51	3.49	0.06	0.29
SIZE	square meter	Housing unit size	97.64	89.25	282.68	14.79	30.63
FLOOR	/	Floor number	12.43	11.00	54.00	1.00	7.96
BEDROOM	/	Number of bed rooms	2.22	2.00	8.00	1.00	0.77
TFLOOR	/	Building height (stories)	23.73	22.00	66.00	3.00	8.19
Location Attributes							
D_CBD	km	Distance to city center	6.95	6.50	36.01	0.26	3.09
ZONE	dummy	33 zones					

Table 2 Estimate results of ps-RS model

Variable	ps-RS Model		
	by_Building	*by_License*	*by_Complex*
Δln(size)	0.972 (6047.610***)	0.983 (6803.740***)	0.999 (8199.110***)
Δln(floor)	0.009 (375.860***)	0.010 (421.530***)	0.008 (395.440***)
Δln(bedroom)	0.005 (65.98***)	0.005 (70.01***)	0.005 (103.81***)
Month Dummy	Yes	Yes	Yes
Adjust_R^2	0.903	0.857	0.813
Obs.	14,394,461	22,281,758	31,636,652

t statistics in parentheses

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

Standard errors clustered by complex.

Table 3 Estimate Result of Hedonic Model

Variables	Coefficient (t-statistic)
ln(SIZE)	1.066 (916.96***)
ln(FLOOR)	0.011 (40.06***)
BEDROOM	0.004 (110.96***)
ZONE Dummies	Yes
Month Dummies	Yes
Intercept	7.65 (915.05***)
Adjusted R^2	0.742
Obs.	444,596

t statistics in parentheses

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

Standard errors clustered by complex.

Table 4: Comparing Index Smoothness:
Three Metrics: Volatility, First-Order Autocorrelation, and Sum of Squared Differences Between Index and its Hodrick-Prescott Representation

	Building-version Ps-RS	License-version Ps-RS	Complex-version Ps-RS	Hedonic
Volatility	0.016	0.023	0.034	0.080
AC(1)	0.599	0.405	0.256	-0.094
sum of the square of deviations of return	0.006	0.011	0.022	0.122