Quantile Regression Method of House Price Index for Newly Building Condominiums in China

Ai Lirong
Wuhan Institute of Shipbuilding Technology
Wuhan 430050, China
e-mail: 3247438@qq.com

He Saiqi, Jin Shengping
School of Science
Wuhan University of Technology
Wuhan 430070, China
e-mail: 1410374478@qq.com
spjin@whut.edu.cn

***Abstract*—The house price index plays a very important role in the real estate economy due to it is an indicator of real estate price changing. Aiming at the compilation model of China's new ordinary residential housing price index based on the matching set of upright adjacent floors, the parameter estimation method is improved, and the ordinary least square regression (OLS) is replaced by quantile regression to eliminate the extreme sensitivity of least square regression to outliers. This paper puts forward the quantitative index to evaluate the advantages and disadvantages of different methods to compile house price index, and makes an empirical analysis of the above ideas based on the loan data of commercial banks. The conclusions we got are more consistent with the actual results, more stable, and in line with the requirements of the third generation of housing price index compilation methods.**

Keywords-house price index; set of upright adjacent floors; sample matching method; quantile regression

I. INTRODUCTION

The house price index is formulated through a certain number of real estate samples and a set of compilation methods based on statistical index theory. It is a relative value used to describe the magnitude and direction of changes in house prices and market prices in a certain area. Scientifically compiling house price index is the basic work of monitoring real estate price changes, and it has important theoretical and practical significance.

The compilation method of the house price index has undergone the first generation methods represented by the median and simple weighted average method, and the second generation methods represented by the sample matching method and the Laspeyres weighting method. The third generation method based on mathematical statistical analysis technology which uses the quality-controlled method has gradually become the development direction of the international housing price index compilation. At present, the two mainstream methods of quality adjustment are the repeat sales model and the hedonic model.

The National Statistics Bureau in China issued the residential sales price indices every month, which divides housing types into two major categories: newly building housing and second-hand housing. As shown in Fig. 1, new building housing is divided into residential and nonresidential housing, residential housing is divided into commercial housing and economic and functional housing, and commercial housing is divided into condominiums and high grade housing. Because of the history reason and house's quanlity, it is fitted to the situation in China at present that newly building housing and second-hand housing are treated and processed separative.

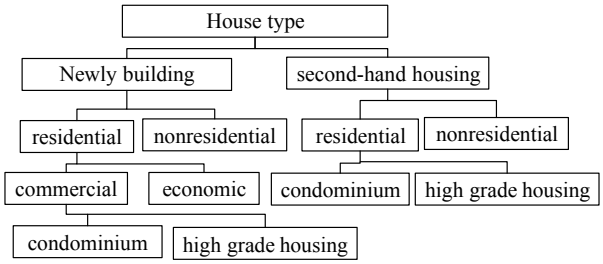

Figure 1. The categories of house type.

Currently, newly building condominiums of most cities in China account for the mainstream housing sales. The state has controlled strictly or prohibited the construction scale of high grade housing. Due to the economic and functional housings which are sold to the lower incomers, the prices are contored by the local goverments and not determined by the market. The prices data from the economic and functional housings should not be used to compile house price index. Indeed the National Statistics Bureau in China does not use the economic and functional housings to issue the residential sales price indices from January, 2018.

On the other hand, the western scholars have established method to compile the house price index for single houses[1]. Although there are condominiums price index such as in Manhattan, New York City, Singapor, Tokyo, these condominiums price index are compiled by using the methods for single houses and the price data for condominiums. They do not make use of the characteristics of the condominiums. So the theory and model to compile the condominiums price index have not been founded[2].

Therefore, it has an important theoretical and practical significance to study the price index compilation method of newly building condominiums in China. We hope that it is easy to achieve breakthrough from the theoretical aspect.

As for the compilation of the housing price index of newly building condominiums in China, most scholars focus on the hedonic method. Since the hedonic method requires the collection of data reflecting more than a dozen or even dozens of characteristic variables of each property, and the characteristic variables of houses may vary greatly in different cities or in different developing stages of the same city. In addition, some characteristic variables are

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00047

difficult to quantify. Therefore, the hedonic method has great limitations both in theory and in practice.

Chinese real estate has its own distinctive features. The newly commercial buildings are all high-rise structures, and most of them are developed in plots of complex. The external neighborhood characteristics of a same real estate or a same complex are basically the same. Intuitively speaking, the structure of newly commercial buildings in China is relatively 'single' or 'homogeneous', which is the core idea of the third generation method of house price index on homogeneity and comparability.

Some progress has been made in the research of the third generation method for compiling the house price index of newly building condominiums in China. Dr. Y. Xu of Xiamen University proposed the repeat-sales-like rule, which skillfully applied the repeat sales model to the new constructed housing market. Guo X. Y., Zheng S. Q., D. Geltner and Liu H. Y. of Tsinghua University proposed the pseudo repeat sales model[3].

However, the team of Xiamen University matched floors based on the property price and did not consider the impact of different orientations on property price, so its matching method may introduce new errors. The team of Tsinghua University used the 'single interval' principle to match, namely, after all samples are sorted by time, only the samples on adjacent time can be matched, which may cause errors due to the large difference in the distribution of properties in two adjacent periods.

McMillen[4] and Deng[5] used propensity scores and the team of Tsinghua University used a method similar to the hedonic method to construct matching pairs. Although the method is general, both of them have the same problem as hedonic price method, i.e. the selection bias of characteristic variables, which results in omitted characteristic variables in practical work and makes the method invalid or inaccurate.

Combined with the characteristics of the newly building condominiums in China, this paper studies the quantile solution method based on our previous work of compiling the house price index model of the newly building condominiums based on the set of upright adjacent floors, and puts forward the evaluation index which is used to compare and analyze the advantages and disadvantages of different methods[6-8].

II. THE MATCHING MODEL TO COMPILE THE HOUSE PRICE INDEX OF NEWLY BUILDING CONDOMINIUMS BASED ON THE SET OF UPRIGHT ADJACENT FLOORS

Urban residents in China are very sensitive to the orientation or location of house properties. For example, the prices of condominiums in the east side are generally much higher than those in the west side of a same building, and prices vary greatly on different floors. Based on the concept of the set of upright adjacent floors which we have put forward, we can deal with the problem that house prices are closely related to orientation or other neighborhood characteristics in the compilation of house price index.

A. The Semilinear Nonparametric Model of House Price Index

The main characteristics of commercial housing are structural characteristics and neighborhood characteristics. The structural characteristics include the floor, the area, the orientation of the house, the total floor of the building and so on. Neighborhood characteristics include factors such as the city area where it is located, and whether the surrounding public facilities are complete. Except for the floor, the other structural characteristics and neighbourhood characteristics of the same set of upright adjacent floors are the same. Table I is the main variables and their descriptions that will be used in the following.

TABLE I. THE MAIN VARIABLES AND THEIR DESCRIPTIONS

Variables	Description
Y_i, y_i	Y_i : Unit sales price of property i , unit: yuan; $y_i = \log(Y_i)$
fl_i, tf_i	fl_i : The floor where property i is located tf_i : The height of the building where property i is located
z_i	Standard height: $z_i = fl_i / tf_i$
t_i	The date (month) corresponding to the sale time of property i
tp_i	Type of building in which the property i is located, multi-storey buildings: $tp_i = 0$, small high-rise buildings: $tp_i = 1$, high-rise buildings: $tp_i = 2$
O_i	A set of other characteristics of property i

The semilinear nonparametric model of the logarithm of the house price is:

$$y_i = \beta_{t_i} + \alpha_1 z_i + \alpha_2 z_i^2 + \gamma_1 tp_i z_i + \gamma_2 tp_i z_i^2 + f(O_i) + \varepsilon_i. \quad (1)$$

Where,

1) β_{t_i} is the fixed effect of time t_i , which is the logarithm of the house price index, i.e. $\beta = (\beta_0, \beta_1, \cdots, \beta_T)^T$ denotes the logarithm of the house price index at all times, and β_0 is the logarithm of the house price index in the base period, let $\beta_0 = 0$.

2) α_1, α_2 are the fixed effect of the standard height and the square of the standard height of the floor, respectively. γ_1, γ_2 are the interaction effect of the standard height and its square with the type of buildings, respectively.

3) $f(O_i)$ denotes the influence of neighbourhood characteristics and other unobserved characteristics of the property i on housing prices.

4) ε_i denotes independent and $N(0, \sigma^2)$ -distributed error terms.

B. The Matching Model of House Price Index Based on Set of Upright Adjacent Floors

Assuming that commercial houses i and j belong to the same set of upright adjacent floors, then $O_i = O_j$, $tp_i = tp_j$, $tf_i = tf_j$. Applying within-pair first differencing based on model (1) will obtain (2):

$$y_i - y_j = \beta_{t_i} - \beta_{t_j} + \alpha_1 (z_i - z_j) + \alpha_2 (z_i^2 - z_j^2) + \gamma_1 tp_i (z_i - z_j) + \gamma_2 tp_i (z_i^2 - z_j^2) + \varepsilon_i - \varepsilon_j. \quad (2)$$

Equation (2) no longer contains function $f(O_i)$, so it does not need to collect various characteristic data of the real estate, and it does not required to consider the form of characteristic function, which solves many problems brought by the hedonic method.

For the distribution of $\varepsilon_i - \varepsilon_j$, we previously assumed that it was independent and identical distribution and had normal distribution, and used OLS to estimate the parameters in (2). This article studies the general situation and uses quantile regression to calculate.

III. QUANTILE REGRESSION OF MATCHED MODEL

The OLS regression method of (2) is based on the mean of the price, while the quantile regression estimates the parameters of any quantile of the price, such as selecting the median, 10% quantile, 90% quantile and so on. In the housing price index preparation, quantile regression is typically applied to the hedonic model, which can examine the changes in the degree of influence of each characteristic variable on the housing price at each quantile, and analyze the main factors for the increase or decrease of housing price. McMillen and Thorsnes(2006) applied quantile regression to the estimation of house price index by repeat sales method, to some extent, to suppress the influence of possible housing renovation or decoration on the overestimation of the housing price index in two transactions[9,10].

The quantile of regression is defined as follows:

$$Q_\tau(y \mid x) = \inf\{y : F(y \mid x) \geq \tau\}.$$

Where $F(y \mid x)$ is the conditional distribution function.

If the quantile regression function is a linear function $Q_\tau(y \mid x) = x^{\mathrm{T}} \beta_\tau$, then the linear quantile regression model is:

$$y_i = x_i^{\mathrm{T}} \beta_\tau + \varepsilon_i, i = 1, \cdots, n. \quad (3)$$

Where any τ -th quantile of ε_i is 0. Then, for any $0 < \tau < 1$, the quantile regression problem can be transformed to solve the following problem:

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^{\mathrm{T}} \beta).$$

Where $\rho_\tau(Z) = Z(\tau - \mathrm{I}(Z < 0))$. When $\tau = 1/2$, it is equivalent to minimize the absolute value of the error, which is the so-called median regression or L_1 norm regression.

The main idea of quantile regression is to modify the objective function of OLS regression, which is very sensitive to outliers. The objective function of quantile regression in model (2) is:

$$\min \sum \rho_\tau \begin{pmatrix} y_i - y_j - \beta_{t_i} + \beta_{t_j} - \alpha_1 (z_i - z_j) \\ -\alpha_2 (z_i^2 - z_j^2) - \gamma_1 tp_i (z_i - z_j) \\ -\gamma_2 tp_i (z_i^2 - z_j^2) \end{pmatrix}. \quad (4)$$

Equation (4) can be solved by simplex method or interior point method of linear programming, or by functions of some software packages, such as rq() in the package 'quant' of R. The determination of the quantile should be based on the sample data and the reasonable range of the housing price change, through a certain number of indicators to evaluate and finally determine the value of the quantile.

IV. EMPIRICAL CALCULATION AND RESULT ANALYSIS

A. Data Source and Calculation Results

The Wuhan Branch of the People's Bank of China collected the mortgage loan data from commercial banks of 83 newly sold commercial houses in Xiangyang from January to December 2012. The transaction price is real and reliable, which can reflect the market situation of new building condominiums in Xiangyang.

After eliminating incomplete information, 6,354 valid samples were obtained. According to (2), the matching process based on upright adjacent floors was carried out, and 4,579 matching pairs were obtained, accounting for 4579/6354=72.1% of the original data.

Using the OLS calculation parameters, the house price index corresponding to the row of 'OLS' in Table II is obtained, where the base index of January is 100. Then, use the quantile regression method to estimate parameters in the model (2) with the data set of all 4,579 matching pairs, and let $\tau = 0.3, 0.5, 0.8$ respectively. The calculation results are shown in the corresponding rows of 'Quant0.3, Quant0.5, Quant0.8' in Table II.

TABLE II. ESTIMATE RESULTS OF OLS METHOD AND QUANTILE REGRESSION

Month / Methods	1	2	3	4
OLS	100.000	100.380	100.488	101.023
Quant0.3	100.000	100.206	100.459	100.375
Quant0.5	**100.000**	**100.016**	**100.063**	**100.157**
Quant0.8	100.000	99.952	100.359	100.639
Month / Methods	**5**	**6**	**7**	**8**
OLS	100.834	101.260	101.632	102.212
Quant0.3	100.633	100.416	100.687	101.116
Quant0.5	**100.427**	**100.323**	**100.442**	**100.701**
Quant0.8	100.449	100.328	100.486	100.850
Month / Methods	**9**	**10**	**11**	**12**
OLS	102.666	103.097	102.833	102.639
Quant0.3	101.444	101.349	101.369	101.313
Quant0.5	**100.794**	**100.806**	**100.810**	**100.810**
Quant0.8	101.145	101.557	101.370	101.459

The housing price index in Table II is drawn as a corresponding line chart, as shown in Fig. 2.

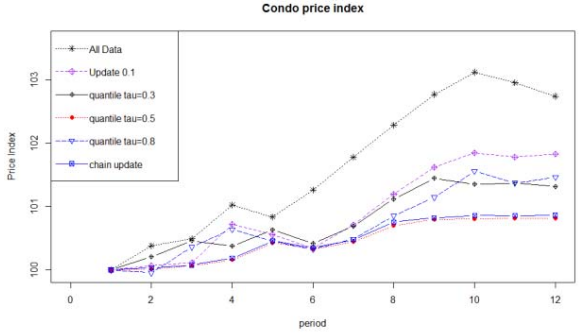

Figure 2. The OLS Index and the Quantile Indices of Xiangyang.

It can be seen from Fig. 2 that the OLS regression and the quantile regression with a quantile far away from 0.5 overestimate the house price index, and the quantile regression with a quantile of 0.5 suppresses matching pairs with too large errors, so the trend of corresponding house price index is relatively flat.

B. The Calculation Results Analysis

In order to compare the advantages and disadvantages of various methods, the following tests are carried out. Randomly select 90% of 4,579 matching pairs as the training set, and the remaining 10% as the test set.

Based on the training set, use OLS method and quantile regression to estimate the parameters in model (2). The estimated parameters are applied to each matching pair (i, j) in the test set, and the predicted value $\hat{y}_i$ of one of the commercial houses can be obtained. This predicted value and its actual price. The difference between the predicted value $\exp(\hat{y}_i) = \hat{Y}_i$ and the true value can be used to test the advantages and disadvantages of the calculation method of the model.

Define the root mean square error($RMSE$) as shown in (5):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}. \quad (5)$$

Where n is the number of matching pairs in the test set, and $Y_i, \hat{Y}_i$ denote the true and predicted house prices of commercial housing i respectively.

Randomly select 457 matching pairs from all data sets as the test set, and the rest as the training set. After a random segmentation, we can get the corresponding $RMSE$ after using OLS and quantile regression to estimate the parameters of (2). Repeat the experiment for 500 times, then calculate the mean value and standard deviation of 500 corresponding $RMSE$ to the two methods respectively, and obtain the corresponding Mean value(Mean(RMSE)), minimum value(Min(RMSE)) and standard deviation(Sd(RMSE)), as shown in Table III:

TABLE III. THE VALUE OF EVALUATION INDEX OF VARIOUS METHODS

Indicators / Methods	Mean Value Mean(RMSE)	Minimum Value Min(RMSE)	Standard Deviation Sd(RMSE)
OLS	292.590	188.140	53.406
Quant0.3	292.011	183.649	52.665
Quant0.5	**291.786**	**184.701**	**52.553**
Quant0.8	293.762	189.771	52.754

Based on the mean, minimum and standard deviation of $RMSE$, the smaller values, the better. Compared with quantile 0.5, the mean, minimum and standard deviation of OLS, quantile 0.3, 0.8 and other quantiles (not listed in Table 3) are relatively larger (except for the minimum value corresponding to quantile 0.3). Thus, the quantile 0.5 is a reasonable choice and an expected conclusion for the calculation of the housing price index.

In Fig. 2, the housing price index corresponding to 'Quant0.5' is smaller than the house price index corresponding to 'All Data' and 'Update0.1', indicating that OLS method greatly overestimates the house price index, mainly because the OLS method is highly sensitive to data with large errors.

V. CONCLUSIONS AND SUGGESTIONS

According to the matching model based on the set of upright adjacent floors proposed by the author's team, this paper will use the quantile regression to study the method of solving the parameters, and put forward quantitative indicators to evaluate the advantages and disadvantages of different compilation methods. The calculation example shows that the quantile regression with quantile 0.5 can reflect the housing price market more stably and truly.

In this paper, the matching model and method based on the set of upright adjacent floors are improved appropriately, which can also be applied to the compilation of the housing price index of the second-hand house. Also, they provide an early theoretical preparation for the compilation of a unified general commercial housing price index when the newly building house and second-hand house market are merged together.

REFERENCES

[1] K. E. Case and R. J. Shiller, "The efficiency of the market for single-family homes," The American Economic Review, vol. 79, Feb. 1988, pp. 125-137, doi:10.3386/w2506.

[2] M. B. David, "S&P CoreLogic Case-Shiller Home Price Indices Methodology," Copyright © 2016 S&P Dow Jones Indices LLC, https://us.spindices.com/index-family/real-estate/sp-corelogic-case-shiller.

[3] Guo X. Y., Zheng S. Q., D. Geltner and Liu H. Y., "A new approach for constructing home price indices: The pseudo repeat sales model and its application in China," Journal of Housing Economics, vol. 25, Sep. 2014, pp. 20-38, doi:10.1016/j.jhe.2014.01.005.

[4] D. P. McMillen, "Repeat sales as a matching estimator," Real Estate Economics, vol. 40, Sep. 2012, pp. 745-773, doi:10.1111/j.1540-6229.2012.00343.x.

[5] Deng Y. , D. P. McMillen and T. F. Sing, "Private residential price indices in Singapore: a matching approach," Regional Science and Urban Economics, vol. 42, May 2012, pp. 485-494, doi:10.1016/j.regsciurbeco.2011.06.004.

[6] Jin S. P., Zeng X. and Li Q., "Research on the model to compile the house price index for newly building condominiums in China," Wuhan Finance, vol. 6, 2017, pp.49-53, doi:CNKI:SUN:YHQY.0.2017-06-012.

[7] Jin W. H. and Jin S. P., "The Virtual repeat sale model for the house price index for new building in China," Applied Mathematics, vol. 5, Dec. 2014 , pp. 3431-3436, doi: 10.4236/am.2014.521320.

[8] Dong Y. L., Jin S. P. and Chen J. Q., "Semiparametric matching model for compiling price index of newly-built ordinary houses," Statistics and Decision, vol. 36, Aug. 2020, pp. 41-44, doi:10.13546/j.cnki.tjyjc.2020.16.009.

[9] D. P. McMillen and T. Paul, "Housing renovations and the quantile repeat‐sales price index," Real Estate Economics, vol. 34, Nov. 2006, pp. 567-584, doi:10.1111/j.1540-6229.2006.00179.x.

[10] Zhang L. and Yi Y., "Quantile condominium price indices in Beijing," Regional Science and Urban Economics, vol. 63, Mar. 2017, pp. 85-96, doi:10.1016/j.regsciurbeco.2017.01.002.