JSPS Grants-in-Aid for Creative Scientific Research

Understanding Inflation Dynamics of the Japanese Economy

Working Paper Series No.56

On the Evolution of the House Price Distribution

Takaaki Ohnishi
Takayuki Mizuno
Chihiro Shimizu
and
Tsutomu Watanabe

April 22, 2010

Research Center for Price Dynamics
Institute of Economic Research, Hitotsubashi University
Naka 2-1, Kunitachi-city, Tokyo 186-8603, JAPAN
Tel/Fax: +81-42-580-9138
E-mail: sousei-sec@ier.hit-u.ac.jp
http://www.ier.hit-u.ac.jp/~ifd/

On the Evolution of the House Price Distribution

Takaaki Ohnishi* Takayuki Mizuno† Chihiro Shimizu‡
Tsutomu Watanabe§

April 22, 2010

Abstract

Is the cross-sectional distribution of house prices close to a (log)normal distribution, as is often assumed in empirical studies on house price indexes? How does it evolve over time? How does it look like during the period of housing bubbles? To address these questions, we investigate the cross-secional distribution of house prices in the Greater Tokyo Area. Using a unique dataset containing individual listings in a widely circulated real estate advertisement magazine in 1986 to 2009, we find the following. First, the house price, P_{it}, is characterized by a distribution with much fatter tails than a lognormal distribution, and the tail part is quite close to that of a power-law or a Pareto distribution. Second, the size of a house, S_i, follows an exponential distribution. These two findings about the distributions of P_{it} and S_i imply that the the price distribution conditional on the house size, i.e., $\Pr(P_{it} \mid S_i)$, follows a lognormal distribution. We confirm this by showing that size adjusted prices indeed follow a lognormal distribution, except for periods of the housing bubble in Tokyo when the price distribution remains asymmetric and skewed to the right even after controlling for the size effect.

JEL Classification Number: R10; C16
Keywords: house prices; house price indexes; power-law distributions; fat tails; hedonic regression; the size dependence of house prices; housing bubbles

1 Introduction

Researches on house prices typically start by producing a time series of the *mean* of prices across housing units in a particular region by, for example, running a hedonic regression or by adopting a repeat-sales method. In this paper, we propose an alternative research strategy: we look at the entire distribution of house prices across housing units in a particular region at a particular point of time, and then investigate the evolution of such cross sectional distributions over time. We seek to describe price dynamics in a housing market not merely by changes in the

*Correspondence: Takaaki Ohnishi, Canon Institute for Global Studies and University of Tokyo. E-mail: ohnishi.takaaki@canon-igs.org. This research is a part of the project entitled: Understanding Inflation Dynamics of the Japanese Economy, funded by JSPS Grant-in-Aid for Creative Scientific Research (18GS0101).
†Hitotsubashi University. E-mail: mizuno@ier.hit-u.ac.jp
‡Reitaku University. E-mail: cshimizu@reitaku-u.ac.jp
§Hitotsubashi University. E-mail: tsutomu.w@srv.cc.hit-u.ac.jp

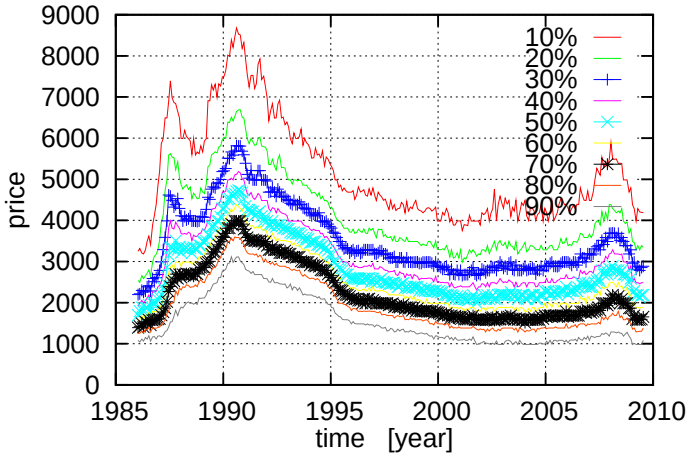

Figure 1: The percentiles of the cross-sectional house price distribution for each month

mean but by changes in some key parameters that fully characterize the entire cross sectional price distribution. Our ultimate goal is to produce a new housing price index based on these key parameters.

Specific questions we have in mind is whether the house price distribution is close to a Gaussian distribution or something else; whether it has fatter-tails than a Gaussian distribution; how the distribution is affected by various attributes of a house, including its size, location, and age; how the shape of the distribution changes over time, especially during the period of bubble and its bursting.

The rest of the paper is organized as follows. Section 2 provides a description of the dataset we will use in this paper. The distributions of house prices and those of house sizes are investigated in sections 3 and 4, respectively. In section 5 we will estimate a size adjusted price for each housing units, and see whether the distribution of the size adjusted prices is close to a normal distribution. Section 6 provides a tentative conclusion.

2 Data

In conducting this empirical exercise, we use a unique dataset that we have compiled from individual listings in a widely circulated real estate advertisement magazine, which is published by Recruit Co., Ltd., one of the largest vendors of residential lettings information in Japan. The dataset covers the Tokyo metropolitan area for the period 1986 to 2009, including the bubble period in the late 1980s and its collapse in the early 90s. It contains 724,416 listings for condominiums and 1,602,918 listings for single family houses. This dataset is used by a series of papers including Shimizu et al (2009) which compares hedonic and repeat-sales measures from various viewpoints. In this paper we will use data only for condominiums.

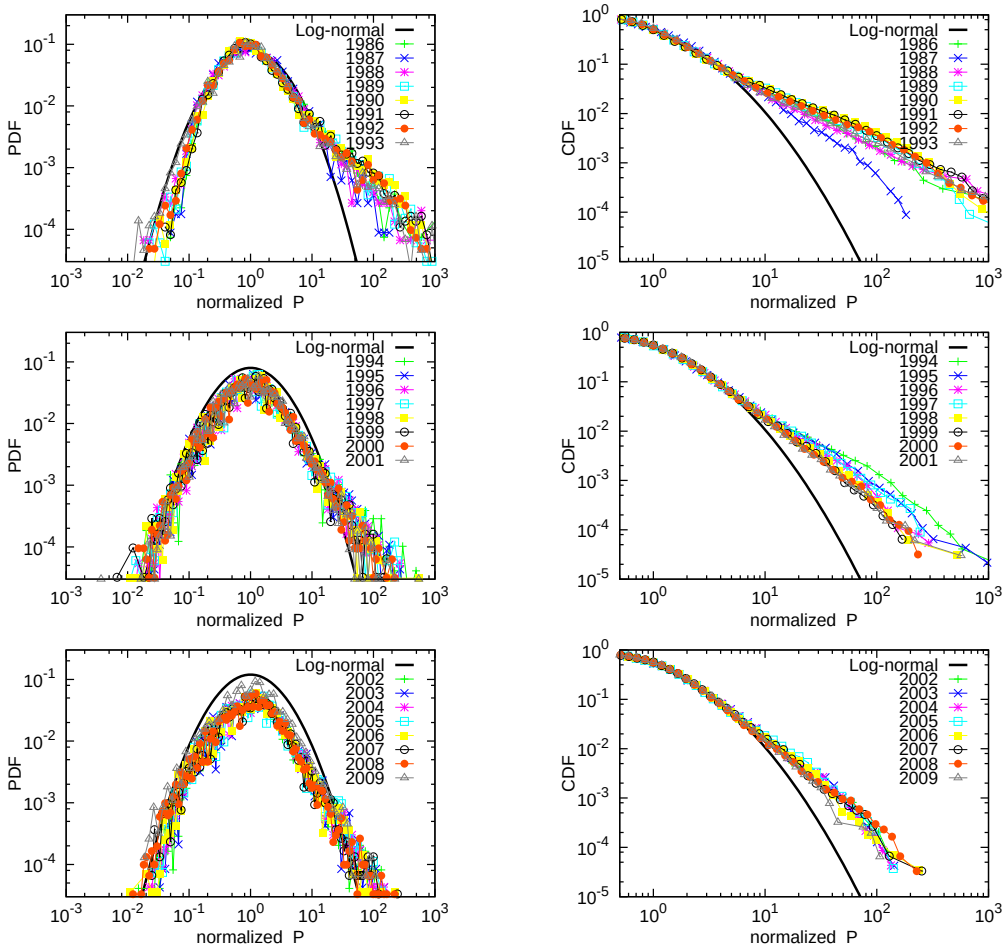

Figure 2: PDFs and CDFs of the house price distribution by year

3 House price distributions

As a first step to look at the data, we show in Fig 1 the monthly evolution of the percentiles of the cross-sectional house price distribution. For example, the 10th percentile, which is shown by the red line, indicates the price level above which 10 percent of the observations may be found. As shown by the 50th percentile (or the median) line, house prices rose rapidly in the latter half of the 1980s and declined in the first half of the 1990s. This swing in prices corresponds to housing bubble and its bursting in Tokyo. We see that the distance between the 90th percentile and the 50th percentile is much smaller than the one between the 10th percentile and the 50th percentile, implying that the distribution is not symmetric. This asymmetry seems to be particularly significant during the bubble period (i.e., the latter half of the 1980s).

To examine more closely the shape of the price distribution, we show in Fig 2 the probability density function (PDF) and the cumulative distribution function (CDF) for each year. Prices

of housing units in each year are normalized using the mean and the standard deviation in that year.[1] Note that the CDFs are constructed by summing up the densities *above* (not below) a particular price level in order to examine closely the right tail (i.e., larger price level) of each distribution. From this figure, we see that the PDFs have much fatter tails than a lognormal distribution, whose PDF and CDF are shown by the solid lines. For example, the fraction of housing units whose prices deviate from the mean by more than 3σ is about 1.71 percent in 2005 while the corresponding figure for a normal distribution is 0.26 percent. More importantly, the deviation from a lognormal distribution tends to be larger in the latter half of the 1980s and the early 1990s; specifically, the PDFs in these years are substantially skewed to the right. This implies that, even during the bubble period, house prices did not rise by an equal percentage for every housing unit, but relative prices among houses changed significantly during those years.

The CDFs in this figure provide much more detailed information regarding how much the price distributions deviate from a lognormal distribution. Specifically, we see that the CDF in each year forms a straight line in this log-log graph, implying that the house price distribution is close to a power-law distribution (or a Pareto distribution) whose PDF and CDF are given by

$$\Pr(P_{it} = p) = \frac{\zeta_t m_t^{\zeta_t}}{p^{\zeta_t+1}}; \qquad \Pr(P_{it} \geq p) = \left(\frac{m_t}{p}\right)^{\zeta_t}; \qquad p > m_t > 0 \tag{1}$$

where P_{it} represents the price of a housing unit i in period t, and ζ_t and m_t are time-variant positive parameters. The CDF given in (1) implies that

$$\ln \Pr(P_{it} \geq p) = -\zeta_t \ln p + \zeta_t \ln m_t$$

In words, the log of the cumulative probability should be linearly related to the log of the price, which is actually observed in Fig 2. The slope of a linear line, i.e. the value of ζ_t, is almost identical across different years and it is about three.[2]

As a goodness-of-fit test, we conduct a test proposed by Malevergne et al. (2009). Specifically, we test the null hypothesis that the house price distribution in each year follows a power-law distribution against the alternative hypothesis that it follows a log normal distribution. We find that the null cannot be rejected for each of the 24 distributions (namely, the distributions from 1986 to 2009) at the five percent significance level.

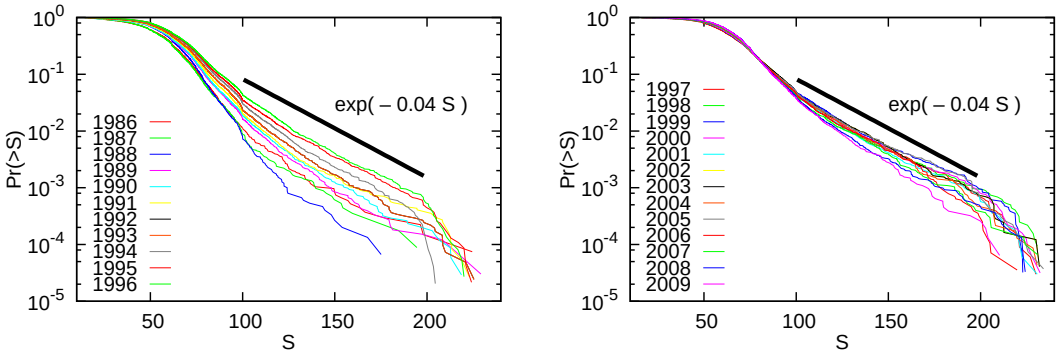

Figure 3: CDFs of the size of a house

4 House size distributions

Previous studies on wealth (or income) distributions across households typically find that those distributions are characterized by fat tails, and that they follow a power-law distribution.[3] Given that a house is an important part of wealth for each household, it may not be so surprising to detect a similar feature in the house price distribution. But why and how (through what mechanism) do house prices follow a power-law distribution? To address this question, we decompose the house price distribution as follows:

$$\Pr(P_{it} = p) = \sum_s \Pr(P_{it} = p \mid S_i = s) \Pr(S_i = s) \tag{2}$$

where S_i represents the size of a housing unit i. The term $\Pr(S_i = s)$ represents the distribution of the size of a house, and the term $\sum \Pr(P_{it} = p \mid S_i = s)$ represents the distribution of the price of a house conditional on its size.

Figure 3 shows the CDFs of the size of a house, measured by square meters, for each year, with the size of a house on the horizontal axis and the log of CDF on the vertical axis. We see that the CDF in each year forms a straight line in this semi-log graph, implying that the size distribution follows an exponential distribution whose PDF and CDF are given by

$$\Pr(S_i = s) = \lambda_t \exp(-\lambda_t s); \qquad \Pr(S_i \geq s) = \exp(-\lambda_t s); \qquad \lambda_t > 0 \tag{3}$$

where λ_t is a time-variant positive parameter. Note that the CDF shown above implies that

$$\ln \Pr(S_i \geq s) = -\lambda_t s$$

[1]Specifically, we define normalized prices as $\exp[(\ln P_{it} - \mu_t)/\sigma_t]$, where μ_t and σ_t are the mean and the standard deviation in year t.

[2]Note that normalized prices are equal to $[\exp(-\mu_t) P_{it}]^{1/\sigma_t}$, so that the slope of each CDF in Fig 2 represents $\sigma_t \zeta_t$ rather than ζ_t. We estimated the value of ζ using the CDFs for the original prices, which are not shown here due to the space limitation.

[3]See Pareto (1896). Gabaix (2009) provides an extensive survey of empirical studies on power laws in various aspects of economic activities, including income and wealth, the size of cities and firms, stock market returns, and so on.

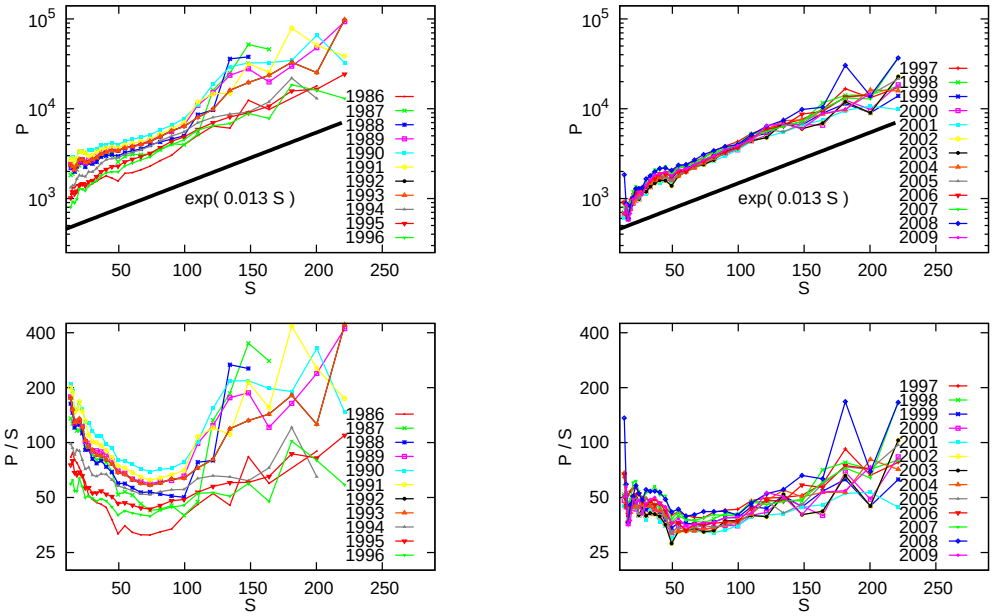

Figure 4: Relationship between the size of a house and its price

so that the log of the CDF depends linearly on the size of a house. The slope of the CDF line, namely the value of λ, is something around 0.04 in each year. The fact that the house size follows an exponential distribution implies that the tails of the size distribution are less fat than the ones of the price distribution; for example, the fraction of housing units whose size deviate from the mean by more than 3σ is only 1.17 percent in 2005, which is smaller than the corresponding figure for the price distribution (1.71 percent) although it is still far greater than the corresponding figure for a normal distribution (0.26 percent).[4]

5 Size-adjusted prices

An important implication of eqns (1) and (3) is that the house price conditional on its size, i.e., $P_{it} \mid S_i = s$ in eq (2), follows a lognormal distribution. Specifically, a size-adjusted price $\tilde{P}_{it}$, which is defined as

$$\tilde{P}_{it} \equiv \frac{P_{it}}{\exp\left(a_t S_i + b_t\right)} \tag{4}$$

where

$$a_t \equiv \frac{\lambda_t}{\zeta_t}; \quad b_t \equiv \ln m_t \tag{5}$$

[4]One may wonder why the house size obeys the exponential distribution. To address this, we set up a simple optimization problem in which one allocates space to each house so as to maximize the variety of house sizes subject to the space constraint (i.e. only limited space is available to be allocated to houses), and show that a solution to this problem is indeed characterized by an exponential distribution.

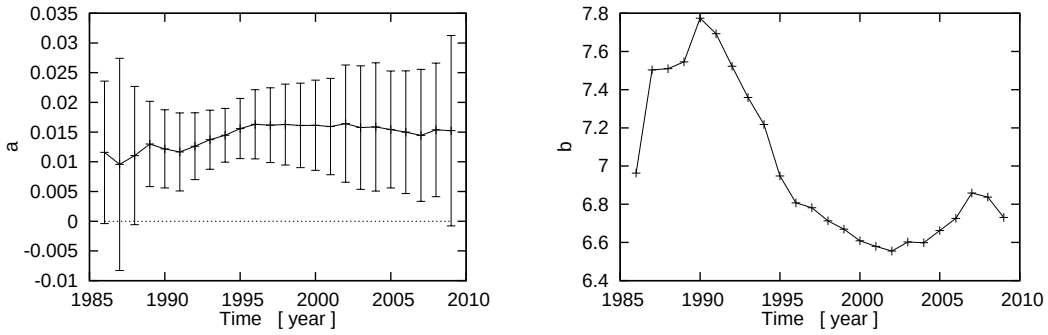

Figure 5: Price-size regressions

follows a lognormal distribution with the mean being unity.[5]

To test this implication, we first examine for a linear relationship between the log of the price of a house and its size. The upper panels of Fig 4, with the house size on the horizontal axis and the median of the log price corresponding to that size on the vertical axis, clearly indicate the presence of such a linear relationship between the two variables. Furthermore, the fact that the size adjusted price defined by eqn (2) follows a lognormal distribution implies that the per unit area price, $P/S = \exp(aS + b)/S$, decreases with S when S is small and increases with S when S is sufficiently large, so that there should exist a U-shaped relationship between the per unit area price and the house size. The lower panels of Fig 4, in which the vertical axis now represents P/S, confirms this prediction.

To give an intuitive understanding of what is going on here, think about a simple example in which the household A has 100 times as much wealth as the household B does, so that the household A spends money for a house 100 times as much as B does. Note that the existence of such a huge difference in wealth and the house price across households is not so rare, given that both wealth and the house price are characterized by distributions with power-law tails. Given this, one may wonder how the A's house looks like. Does it have a bathroom that is 100 times larger than the one in the B's house? Alternatively, does it have 100 bathrooms? Needless to say, neither is true; because such a giant bathroom (or so many bathrooms) is nothing but uncomfortable even to millionaire like A. Instead, the size of the A's house (and therefore the size of its bathroom) is probably no more than 10 times, as implied by the fact that the house size follows a distribution with much less fat tails compared to the house price. To fill the gap, the unit area price for the A's house must be 10 times higher.

We now proceed to estimating a_t and b_t in eq (4) by running a regression of the form

$$\ln P_{it} = a_t S_i + b_t + \epsilon_{it} \tag{6}$$

[5]The price-size relationship described by eqn (4) provides an answer to the question regarding the choice of functional form for hedonic price equations, which has been extensively discussed by previous studies, such as Cropper et al (1988), Diewert (2003), and Triplett (2004). The novelty of our approach is that we derive this functional form not from economic theories but from the statistical fact that the price and the size of a house obeys a power-law distribution and an exponential distribution, respectively.

where ϵ_{it} is a disturbance term obeying a normal distribution with the zero mean. The result is presented in Fig 5. The estimate of a_t in each year is around 0.015, implying that an increase in the house size by a square meter leads to a 1.5 percent increase in the house price. We see that the estimates of a_t are almost identical across years in the sense that the changes in a_t across years are within the confidence intervals. More importantly, the estimate of a is sufficiently close to the value implied by eqn (5). That is, the slope of the CDF line in the house price in Fig 2 (i.e., the value of ζ) is around 3, and the slope of the CDF line in the house size in Fig 3 (i.e., the value of λ) is equal to 0.04, so that $\lambda/\zeta \approx 0.013$. This number is quite close to the point estimate of a in each year, and is within the confidence intervals.[6] Turning to the estimate of b_t, it exhibits substantial fluctuations: it increased by more than 20 percent per year from 1986 to 1990, and it declined by 10 percent per year from 1990 to 2002.

Finally, we construct the size adjusted prices $\tilde{P}_{it}$ by using the estimates of a and b in order to see whether it does indeed follow a lognormal distribution. Specifically, we assume that a_t is identical across years and that it equals the sample average ($\hat{a} = 0.0125$). Based on this, we calculate $P_{it}/\exp(\hat{a}S_i)$, whose CDFs are shown on the right hand side of Fig 6. The price distributions *without* size adjustments (the same figures as in Fig 2) are shown on the left hand side. Comparing these two sets of CDFs, we see that the CDFs for the size adjusted prices are much closer to the CDF of a lognormal distribution, which is shown by the black solid line. Specifically, we see that the CDFs for 2002 to 2009, which are shown on the bottom right panel, are almost identical to the CDF of a lognormal distribution. The same thing applies to the CDFs for 1995 to 2001, which are shown on the middle right panel. However, the CDFs for 1986 to 1994, which are presented on the top right panel, are still far from the CDF of a lognormal distribution, although they are somewhat closer to it as compared to the CDFs of the non-adjusted prices. To measure the distance between the distribution of the adjusted price and a lognormal distribution in a more formal way, we present in Fig 7 quantile-quantile plots of the log of the size-adjusted price against a normal distribution. We see that the dots are on the 45 degree line for the latter half of the sample period (i.e. 1997-2009), indicating that the two distributions are sufficiently close to each other. However, for the former half of the sample period (1986-1996), the dots deviate considerably from the 45 degree line, which clearly rejects the null that the two distributions are identical.

In sum, the results presented in figures 6 and 7 indicate that size adjusted prices follow a lognormal distribution for the quiet periods without extremely large fluctuations in prices. On the other hand, for the periods with extremely large price fluctuations, such as those caused by housing bubbles, the fat tails of price distribution remain largely unchanged even after controlling for the size effect. An implication of these two different results is that we may be able to use the distance of the size adjusted price distribution from a lognormal distribution as a measure of deviations from fundamental price levels. That is, we may be able to say that

[6]Note that the per unit area price, $\exp(aS+b)/S$ takes a minimum value when S is equal to $1/a$. Given the estimate of a ($a = 0.013$), this implies that the per unit area price takes a minimum value when $S = 1/0.013 \approx 75$, which is consistent with what we see in the lower two panels in Fig 4.

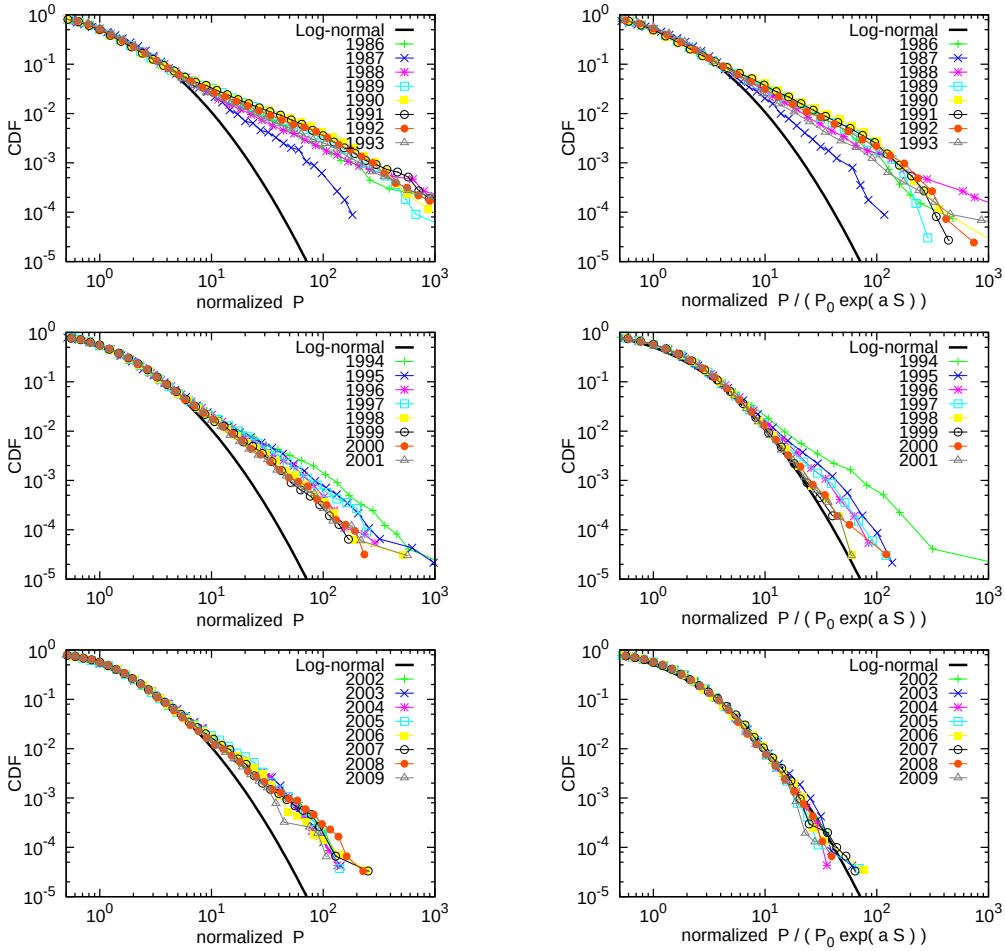

Figure 6: CDFs of the size adjusted house price distribution

prices are sufficiently close to fundamental levels when size adjusted prices follow a lognormal distribution. On the other hand, when the distance between the distribution of size adjusted prices and that of a lognormal distribution is not trivial, then we may be able to say that prices deviate from fundamental values, and in that sense there exist price bubbles. Of course, we need to conduct further investigations before interpreting the results in figures 6 and 7 in this manner.[7] However, aggregate measures of housing prices, either estimated by hedonic regressions or repeat-sales methods, typically focus on the mean of the price distribution, thereby discarding other information on the distribution, including its variance, skewness, and so on. It would be quite difficult (or maybe impossible) to discriminate between price bubbles

[7]Specifically, we may identify housing units that are located in the thick tail of the distribution in Fig 6. These housing units are outliers in the sense that they are unlikely to be observed if the distribution is close to a lognormal. We may be able to investigate whether these housing units are affected by price bubbles by looking at, for example, whether they are located in a particular area with high turnover. This is the task we are currently working on.

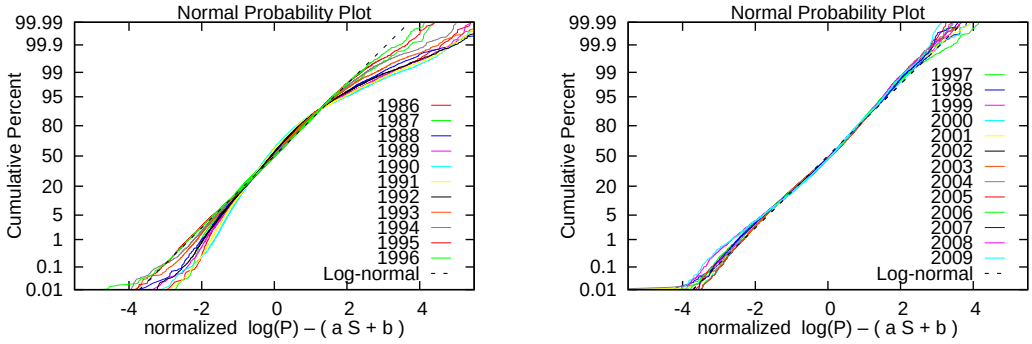

Figure 7: Quantile-quantile plots of the size-adjusted house price distribution against a normal distribution

and changes in fundamental values, if one sticks to these aggregate measures. In this sense our results may suggest the possibility of detecting price bubbles by making full use of information on the entire price distribution.

6 Tentative conclusion

The main findings of this paper can be summarized as follows. First, we have found that the house price for the housing unit i in year t, P_{it}, is characterized by a distribution with much fatter tails than a lognormal distribution, and the tail part is quite close to that of a power-law or a Pareto distribution. Second, we have found that the size of a house, S_i, follows an exponential distribution. An important implication of these two findings is that the price distribution conditional on the house size, i.e., $\Pr(P_{it} \mid S_i)$, is characterized by a lognormal distribution. We confirm this implication by showing that size adjusted prices indeed follow a lognormal distribution, except for periods of the housing bubble in Tokyo when the price distribution remains asymmetric and skewed to the right even after controlling for the size effect.

The fact that size-adjusted prices in each year follow a lognormal distribution implies that one can characterize the evolution of the house price distribution only by two parameters: namely, the mean and the variance of the lognormal distribution in each year. This suggests a new approach to constructing a housing price index. This is similar to the well know approach using hedonic functions in the sense that both are based on the idea that price differences stem from differences in the attributes of a house, but our approach differs from the hedonic approach in some important respects. First, we *derive* a functional form of the price-size relationship only from the statistical fact that the price and the size of a house follows, respectively, a power-law and an exponential distribution. We do not need to rely on any economic theories: we do not need to compare hundreds of regression results with different functional forms (e.g. with or without log; log-log; semi-log; and so on). Second, we have a clearer criteria and procedure to decide how many and which attributes of a house should be considered. Our procedure is quite

simple; we just conduct a goodness-of-fit test to make sure that the house price distribution is sufficiently close to a lognormal distribution after controlling for some attributes of a house.[8] Third, we do not need to conduct regressions in estimating the coefficient on the size of a house in the price equation. Instead, we estimate the slopes of the CDFs for the price and the size distributions in each year, which correspond to the exponents of the power-law and the exponential distributions (ζ_t and λ_t). Then we just calculate the ratio of the two, λ_t/ζ_t, to obtain an estimate for the coefficient on the house size in the price equation. Since we do not rely on regressions, we do not have to worry about various assumptions needed in conducting regressions. Also, we do not pool data for several periods, as is often done in hedonic regressions; all we need is the cross-sectional data in a particular year.

References

[1] Cropper, M. L., L. B. Deck, and K. E. McConnell (1988), "On the Choice of Functional Form for Hedonic Price Functions," Review of Economics and Statistics, Vol. 70, No. 4., 668-675.

[2] Diewert, W. Erwin. (2003), "Hedonic Regressions: A Consumer Theory Approach," in R. C. Feenstra and M. D. Shapiro (eds.), *Scanner Data and Price Indexes*, National Bureau of Economic Research Studies in Income and Wealth, Vol. 64. Chicago, IL: University of Chicago Press, 317-48.

[3] Diewert, W. Erwin, Jan de Haan, and Rens Hendriks (2010), "The Decomposition of a House Price index into Land and Structures Components: A Hedonic Regression Approach," Discussion Paper 10-01, University of British Columbia.

[4] Gabaix, X., (2009), "Power Laws in Economics and Finance," New York University.

[5] Malevergne, Y., V. Pisarenko, and D. Sornette (2009), "Gibrat's law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal," American Economic Review, forthcoming.

[6] Shimizu, C., K. G. Nishimura, and T. Watanabe (2009), "House Prices in Tokyo: A Comparison of Repeat-Sales and Hedonic Measures," Research Center for Price Dynamics Discussion Paper No. 48, November 2009.

[7] Triplett, J. (2004), *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*, OECD Science, Technology and Industry Working Papers 2004/9.

[8] In this paper, we have found that the size adjustment makes the price distribution sufficiently close to a lognormal distribution at least in years without large fluctuations in prices. However, as argued by Diewert et al. (2010) and others, the size and the location of a house are the two most important attributes to determine its price. We plan to see whether the distance between the price distribution and a lognormal distribution would be reduced significantly by considering the location of a house (e.g., commuting time to the central business district) as an additional attribute.