

House prices from magazines, realtors, and the Land Registry

Chihiro Shimizu, Kiyohiko G Nishimura and Tsutomu Watanabe¹

I. Introduction

In constructing a housing price index, one has to make several non-trivial choices. One of them is the choice among alternative estimation methods, such as repeat-sales regression, hedonic regression, etc. There are numerous papers on this issue, both theoretical and empirical. Shimizu et al. (2010), for example, conduct a statistical comparison of several alternative estimation methods using Japanese data. However, there is another important issue which has not been discussed much in the literature, but which has been regarded as critically important from a practical viewpoint: the choice among different data sources for housing prices. There are several types of datasets for housing prices: datasets collected by real estate agencies and associations; datasets provided by mortgage lenders; datasets provided by government departments or institutions; and datasets gathered and provided by newspapers, magazines, and websites.² Needless to say, different datasets contain different types of prices, including sellers' asking prices, transaction prices, valuation prices, etc.

With multiple datasets available, one may ask several questions. Are these prices different? If so, how do they differ from one another? Given the specific purpose of the housing price index one seeks to construct, which dataset is the most suitable? Alternatively, with only one dataset available in a particular country, one may ask whether this is suitable for the purpose of the index one seeks to construct. This paper is a first attempt to address some of these questions.

Specifically, in order to do so, we will conduct a statistical comparison of different house prices collected at different stages of the house buying/selling process. To conduct this exercise, we focus on four different types of prices: (1) asking prices at which properties are initially listed in a magazine, (2) asking prices when an offer for a property is eventually made and the listing is removed from the magazine, (3) contract prices reported by realtors after mortgage approval, and (4) registry prices. We prepare datasets of these four prices for condominiums traded in the Greater Tokyo Area from September 2005 to December 2009. The four prices are collected by different institutions and therefore recorded in different datasets: (1) and (2) are collected by a real estate advertisement magazine; (3) is collected by an association of real estate agents; and (4) is collected jointly by the Land Registry and the Ministry of Land, Infrastructure, Transport and Tourism.

¹ Chihiro Shimizu is at Reitaku University. Kiyohiko G Nishimura is at the Bank of Japan. Tsutomu Watanabe is at the University of Tokyo. This is a shortened version of "House prices at different stages of the buying/selling process". We would like to thank Yongheng Deng, Erwin Diewert, David Fenwick, Sadao Sakamoto, and Hiwon Yoon for helpful discussions and comments. Nishimura's contribution was made mostly before he joined the Policy Board of the Bank of Japan.

² Eurostat (2011) provides a summary of the sources of price information in various countries. For example, in Bulgaria, Canada, the Czech Republic, Estonia, France, Ireland, Latvia, Luxembourg, Poland, Spain and the United States, price data collected by statistical institutes or ministries is used. In Denmark, Finland, Hong Kong SAR, Lithuania, the Netherlands, Norway, Slovenia, Sweden and the United Kingdom, information gathered for registration or taxation purposes is used. In Belgium, France, Germany, Greece, Italy, Portugal and Slovakia, data from real estate agents and associations, research institutes or property consultancies is used. Finally, in Austria, Hungary, Malta and Romania, data from newspapers or websites are used.

An important advantage of prices at earlier stages of the house buying/selling process, such as initial asking prices in a magazine, is that they are likely to be available earlier, so that house price indices based on these prices become available in a timely manner. The issue of timeliness is important given that it takes more than 30 weeks before registry prices become available. On the other hand, it is often said that prices at different stages of the buying/selling process behave quite differently. For example, it is said that when the housing market is, say, in a downturn, prices at earlier stages of the buying/selling process, such as initial asking prices, will tend to be higher than prices at later stages. Also, it is said that, for various reasons, prices at earlier stages contain non-negligible amounts of “noise.” For instance, prices can be renegotiated extensively before a deal is finalised, and not all of the prices appearing at earlier stages end in transactions, for example, because a potential buyer's mortgage application is not approved.

The main question of this paper is whether the four prices differ from each other and, if so, by how much. We will focus on the entire cross-sectional distribution for each of the four prices to make a judgment on whether the four prices are different or not. The cross-sectional distributions for the four prices may differ from each other simply because the datasets in which they are recorded contain houses with different characteristics. For example, the dataset from the magazine may contain more houses with a small floor space than the registry dataset, which may give rise to different price distributions. Therefore, the key to our exercise is how to eliminate quality differences before comparing price distributions. We will conduct quality adjustments in two different ways. The first is to use only the intersection of two different datasets, that is, observations that appear in two datasets. For example, when testing whether initial asking prices in the magazine have a distribution similar to that of registry prices, we first identify houses that appear in both the magazine dataset and the registry dataset and then compare the price distributions for those houses in both datasets. The second method is based on hedonic regressions.

II. Data

We focus on the prices of condominiums traded in the Greater Tokyo Area from September 2005 to December 2009. According to the register information published by the Legal Affairs Bureau, the total number of transactions for condominiums carried out in the Greater Tokyo Area during this period was 360,243. Ideally, we would like to have price information for this entire “universe,” but all we can observe is part of this universe. Specifically, we have three different datasets, each of which is sampled from this universe.

The first is the dataset collected by a weekly magazine, *Shukan Jutaku Joho* (Residential Information Weekly) published by Recruit Co., Ltd., one of the largest vendors of residential lettings information in Japan. This dataset contains initial asking prices (i.e., the asking prices initially set by sellers), denoted by P_1 , and final asking prices (i.e., asking prices immediately before they were removed from the magazine because potential buyers had made an offer), denoted by P_2 . The number of observations for P_1 and P_2 is 155,347, meaning that this dataset covers 43 per cent of the universe. There may exist differences between P_1 and P_2 for various reasons. For example, if the housing market is in a downturn, a seller may have to lower the price to attract buyers. Then P_2 will be lower than P_1 . If the market is very weak, it is even possible that a seller may give up trying to sell the house and thus withdraw it from the market. If this is the case, P_1 is recorded but P_2 is not.

The second dataset is a dataset collected by an association of real estate agents. This dataset is compiled and updated through the Real Estate Information Network System, or REINS, a data network that was developed using multiple listing services in the United States and Canada as a model. This dataset contains transaction prices at the time when the actual sales contracts are made, after the approval of any mortgages. They are denoted by

P_3 . Each price in the dataset is reported by the real estate agent who is involved in the transaction as a broker. The number of observations is 122,547, for a coverage of 34 per cent. Note that P_3 may be different from P_2 because a seller and a buyer may renegotiate the price even after the listing is removed from the magazine. It is possible that P_3 for a particular house may not be recorded in the realtor dataset although P_2 for that house is recorded in the magazine dataset. Specifically, there are more than a few cases where the sale was not successfully concluded because a mortgage application was turned down after the listing had been removed from the magazine.

The third dataset is compiled by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT). We refer to this price as P_4 . In Japan, each transaction must be registered with the Legal Affairs Bureau, but the registered information does not contain transaction prices. To find out transaction prices, the MLIT sends a questionnaire to buyers to collect price information. The number of observations contained in this registry dataset is 58,949, for a coverage of 16 per cent. Since P_3 and P_4 are both transaction prices, there is no clear institutional reason for any discrepancy between the two prices; however, it is still possible for these two prices to differ, partly because they are reported by different parties: a real estate agent for P_3 and the buyer for P_4 . There may be reporting mistakes, intentional and unintentional, on the side of real estate agents, or on the side of buyers, or on both sides.

Some housing units appear in only one of the three datasets, but others appear in two or three datasets. Using address information, we identify those housing units which appear in two or all three of the datasets. For example, the number of housing units that appear both in the magazine dataset and in the registry dataset is 15,015; the number of housing units that are in the magazine dataset but not in the registry dataset is 140,332; and the number of housing units that are in the registry dataset but not in the magazine dataset is 43,934. This clearly indicates that these two datasets contain a large number of different housing units, implying that the statistical properties of the two datasets may be substantially different. This suggests that it may be possible for the three datasets to produce three different house price indices, which behave quite differently, even if the identical estimation method is applied to each of the three datasets.

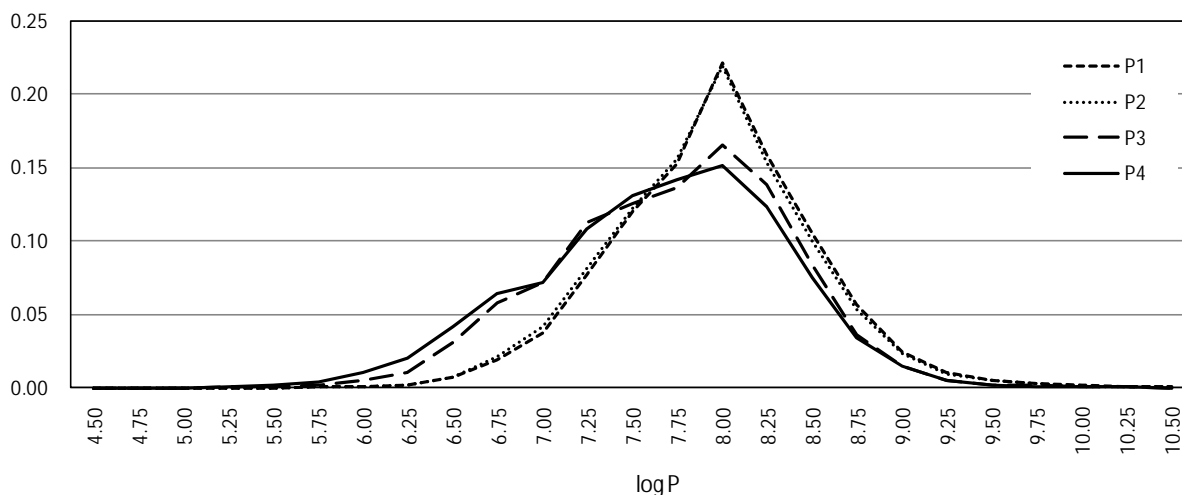
III. Four prices at different stages of the house buying/selling process

Figure 1 shows the cross-sectional distributions for the log of the four prices. The horizontal axis represents the log price while the vertical axis represents the corresponding density. We see that the distributions of P_1 and P_2 are quite similar to each other. On the other hand, the distribution of P_3 differs substantially from the distribution of P_2 ; namely, the distribution of P_2 is almost symmetric, while the distribution of P_3 has a thicker lower tail, implying that the sample of P_3 contains more low-priced houses than the sample of P_2 . This difference in the two distributions may be a reflection of differences in prices at different stages of the house buying/selling process, but it is also possible that the difference in the price distributions may come from differences in the characteristics of the houses in the two datasets.

To investigate this in more detail, we compare the distributions of house attributes for each of the three datasets. The top panel of Figure 2 shows the distributions of floor space, measured in square metres, for the three datasets. The distribution labelled " P_1 and P_2 ," which is from the magazine dataset, is almost symmetric, while the distribution labelled " P_3 ," which is from the realtor dataset, has a thicker lower tail, indicating that the realtor dataset contains more small-sized houses whose floor space is 30 square metres or less. This pattern is even more pronounced in the registry dataset, i.e., the distribution labelled " P_4 ". Turning to the middle and bottom panels of Figure 2, we see that there are substantial

differences between the three datasets in terms of the age of buildings and the distance to the nearest station.

Figure 1
Price densities for P_1 , P_2 , P_3 and P_4



These differences in the distributions of house attributes may be related to the differences in the distributions of house prices. More specifically, the different price distributions we saw in Figure 1 may be mainly due to differences in the composition of houses in terms of their size, age, location, etc. Put differently, it could be that the price distributions are identical once quality differences are controlled for in an appropriate manner.

We will conduct quality adjustments by using only the intersection of two different datasets, that is, observations that appear in two datasets. For example, when testing whether initial asking prices in the magazine have a distribution similar to that of registry prices, we first identify houses that appear in both the magazine dataset and the registry dataset and then compare the price distributions for those houses in both datasets. In this way, we ensure that the two price distributions are not affected by differences in house attributes between the two datasets. This idea is quite similar to the one adopted in the repeat sales method, which is extensively used in constructing quality-adjusted house price indices. As is often pointed out, however, repeat sales samples may not necessarily be representative because houses that are traded multiple times may have certain characteristics that make them different from other houses. A similar type of sample selection bias may arise even in our intersection approach. Houses in the intersection of the magazine dataset and the registry dataset are cases which successfully ended in a transaction. Put differently, houses whose initial asking prices were listed in the magazine but which failed to get an offer from buyers, or where potential buyers failed to get approval for a mortgage, are not included in the intersection.³

IV. Results

The magazine dataset, which contains P_1 and P_2 , and the registry dataset, which contains P_4 , have 15,015 observations in common. On the other hand, there are 22,613 observations

³ See Shimizu et al. (2011) for an alternative way to conduct quality adjustment.

in the intersection of the realtor dataset, which contains P_3 , and the registry dataset, which contains P_4 . We will use these two intersection samples to estimate the distance between the distributions of prices at different stages of the house buying/selling process.

Figure 2
Density functions for house attributes

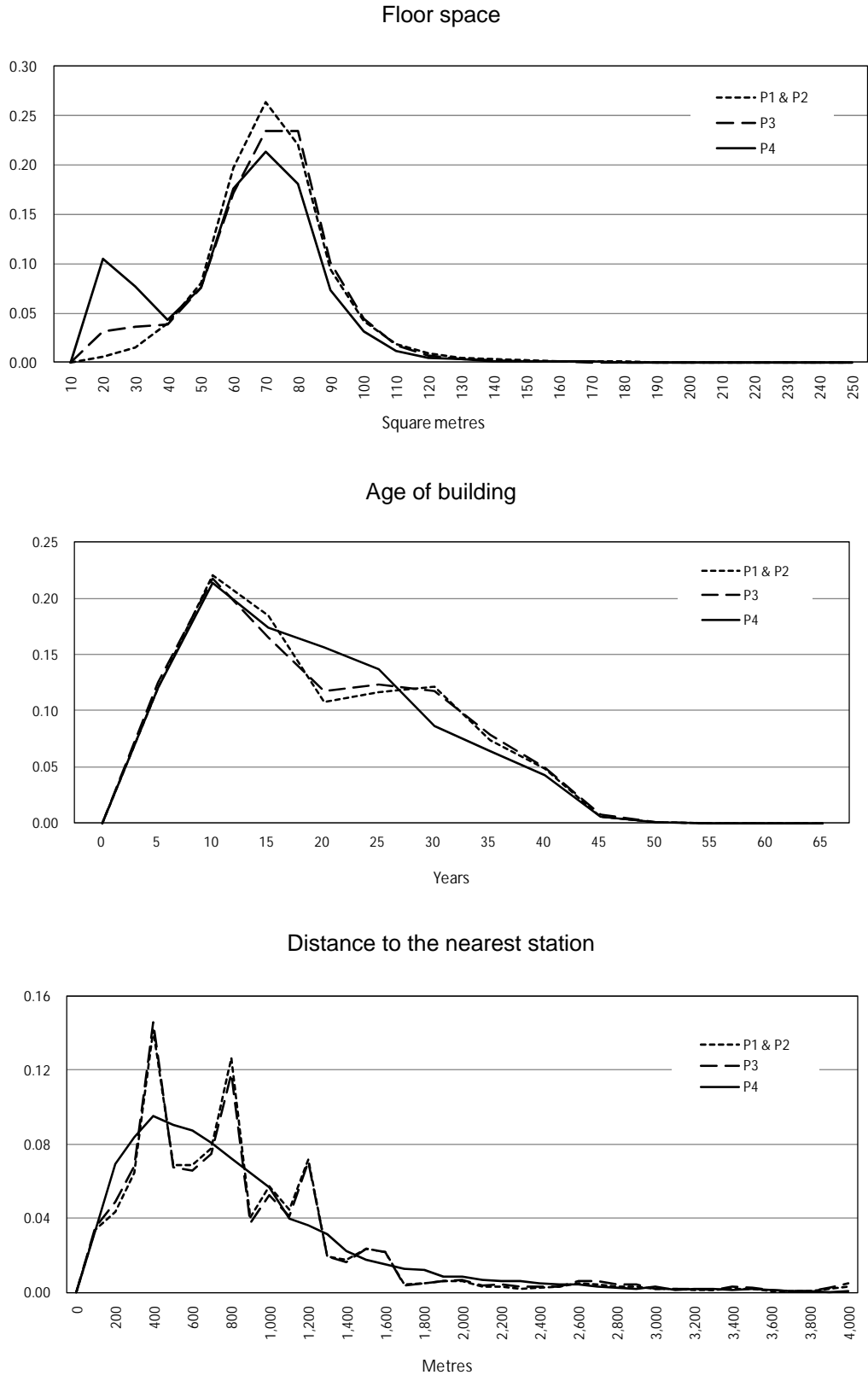


Figure 3 shows the distribution of prices using the intersection samples. The top panel compares the distributions of P_1 and P_4 using the intersection sample of the magazine and registry datasets. In Figure 1, we saw that the distributions of P_1 and P_4 are quite different. However, we now find that the difference between the two distributions is much smaller than before, clearly showing the importance of adjusting for quality. However, the two distributions are not exactly identical even after the quality adjustment. Specifically, the distribution of P_4 has a thicker lower tail than the distribution of P_1 . This may be interpreted as reflecting the fact that asking prices initially listed in the magazine were revised downward during the house selling/purchase process.

The middle panel in Figure 3 compares the distributions of P_2 and P_4 using the intersection sample of the magazine and registry datasets, while the bottom panel compares the distributions of P_3 and P_4 using the intersection sample of the realtor and registry datasets. Both panels show that the differences between the distributions are much smaller than we saw in Figure 1, but there still remain some differences.

In order to see how close the distributions of the four prices are, we draw quantile-quantile (q-q) plots, which provide a graphical technique for determining if two datasets come from populations with a common distribution. The q-q plots are shown in Figure 4, where the quantiles of the first set of prices are plotted against the quantiles of the second set of prices. If the two sets of prices come from populations with the same distribution, the dots should fall along the 45-degree reference line. The greater the departure from this reference line is, the more this suggests that the two sets of prices come from populations with different distributions.

The panels in Figure 4(a) show the q-q plots for raw prices, the distributions of which were shown in Figure 1. The top panel shows the result for P_1 and P_4 , with the log of P_4 on the horizontal axis and the log of P_1 on the vertical axis. Similarly, the middle and bottom panels show the results for P_2 and P_4 and for P_3 and P_4 . The three panels all show that the dots are not exactly on the 45-degree line. For example, in the top panel, the dots are above the 45-degree line; moreover, they deviate more from the 45-degree line for low price ranges, indicating that the distribution of P_4 has a thicker lower tail than that of P_1 . A similar deviation from the 45-degree line can be seen in the q-q plot for P_2 vs. P_4 and the q-q plot for P_3 vs. P_4 , although the deviation is smaller in the case of P_3 vs. P_4 than in the other two cases.

Turning to the q-q plots for quality-adjusted prices by the intersection approach, which are presented in Figure 4(b), we see that the dots are much closer to the 45-degree line than before, although there still remains some deviation from the 45-degree line.

V. Conclusion

In constructing a housing price index, one has to make at least two important choices. The first is the choice among alternative estimation methods. The second is the choice among different data sources for house prices. The choice of the dataset has been regarded as critically important from the practical viewpoint, but has not been discussed much in the literature. This study sought to fill this gap by comparing the distribution of prices collected at different stages of the house buying/selling process, including (1) asking prices at which properties are initially listed in a magazine, (2) asking prices when an offer is eventually made, (3) contract prices reported by realtors, and (4) registry prices. These four prices are collected by different parties and recorded in different datasets. We found that there exist substantial differences between the distributions of the four prices, as well as between the distributions of house attributes. However, once quality differences are controlled for, there remain only small differences between the price distributions. This suggests that prices collected at different stages of the house buying/selling process are still comparable, and

therefore useful in constructing a house price index, as long as they are quality-adjusted in an appropriate manner.

Figure 3
Price densities for housing units observed in two datasets

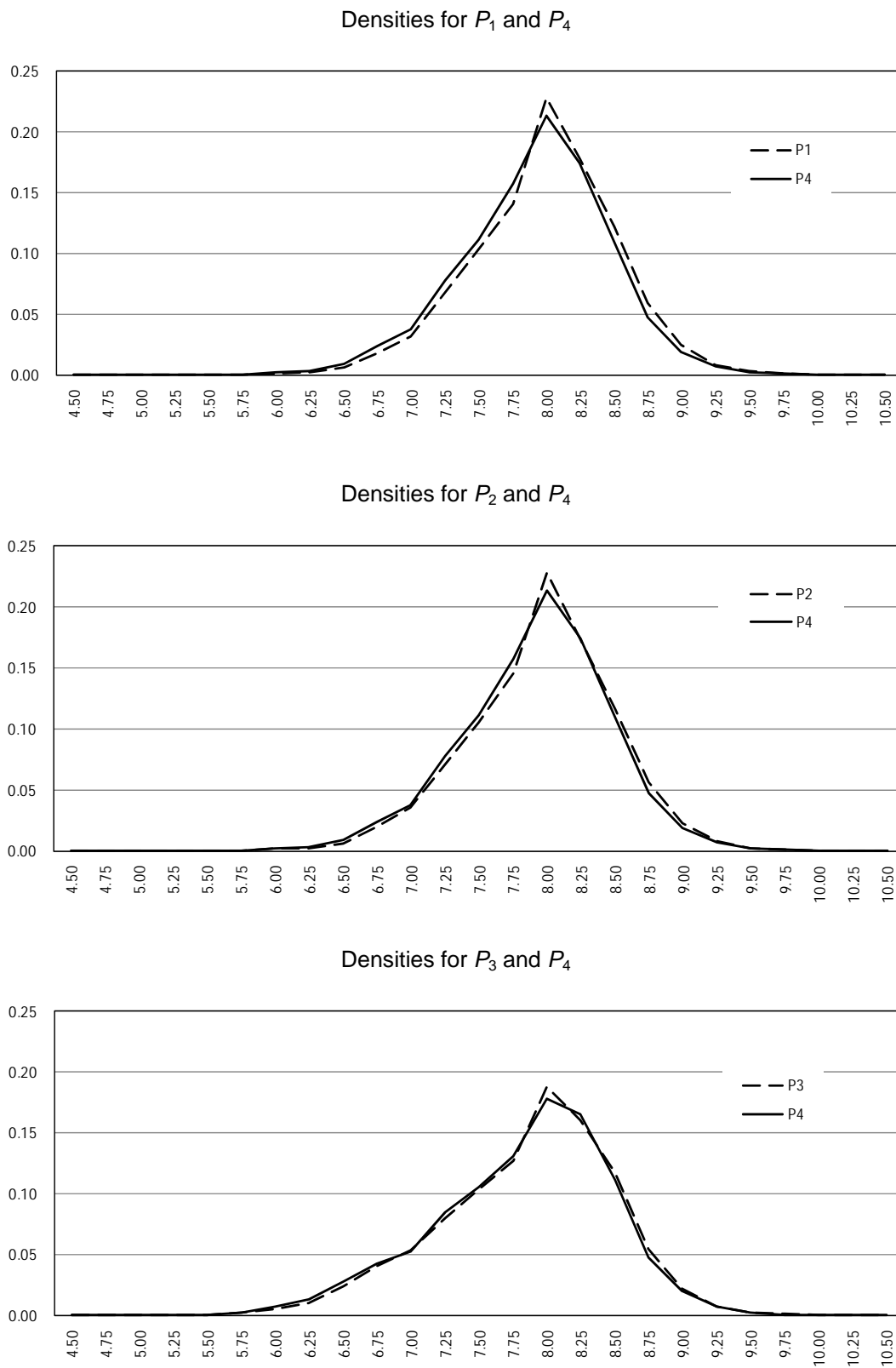


Figure 4(a)
Quantile-quantile plots for raw prices

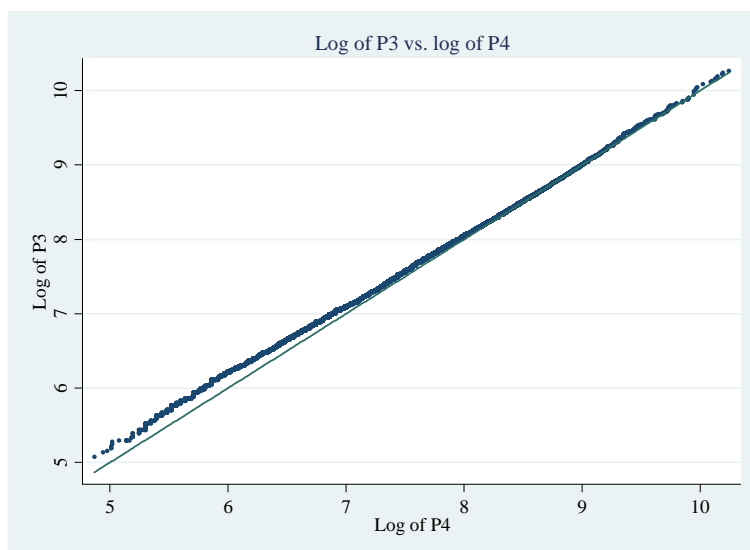
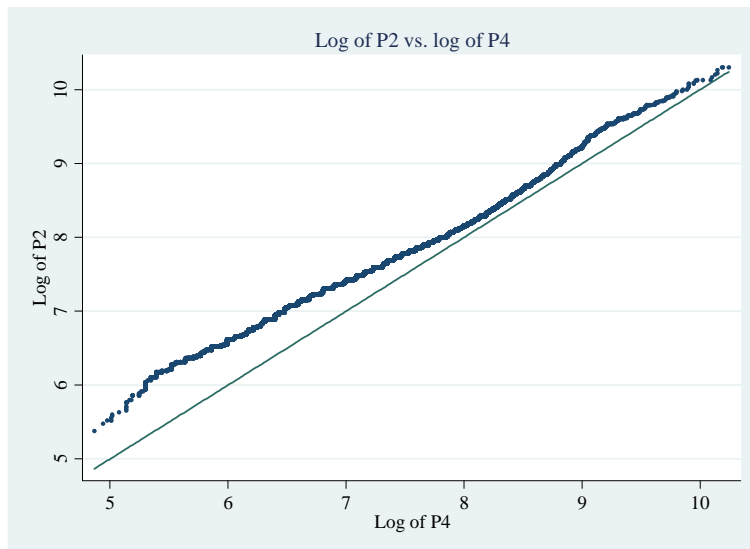
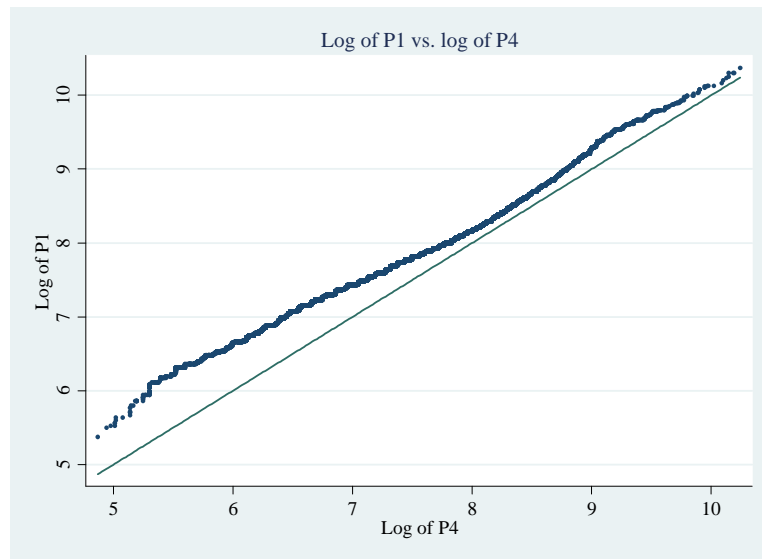
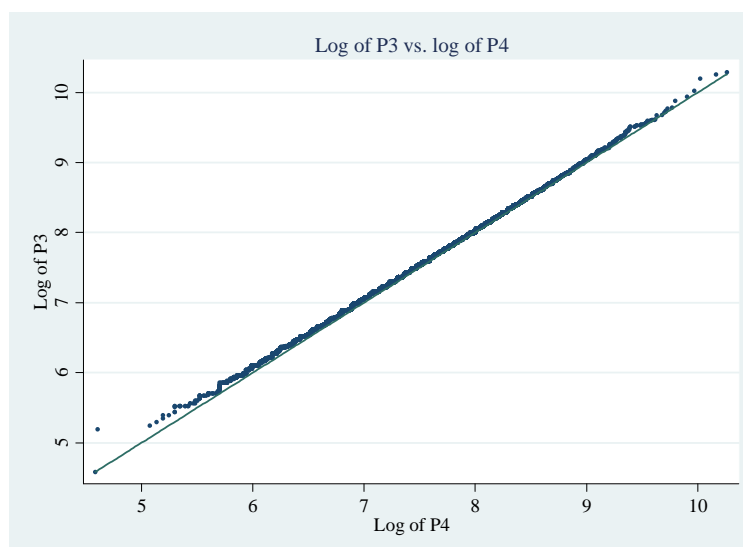
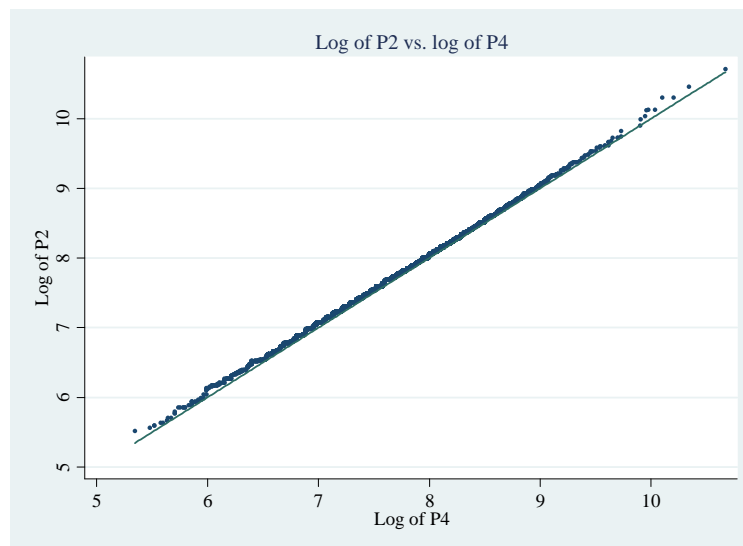
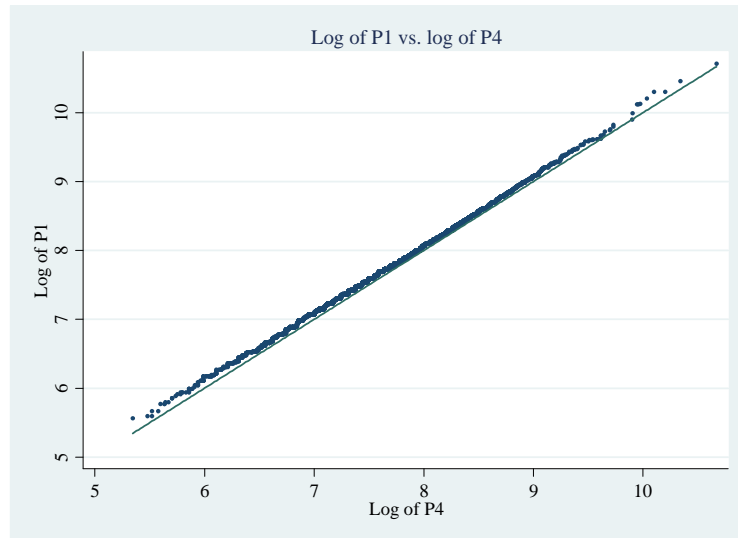


Figure 4(b)

Quantile-quantile plots for quality-adjusted prices by intersection approach



References

Eurostat (2011), *Handbook on residential property price indices*, March 2011. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/hicp/methodology/owner_occupied_housing_hpi/rppi_handbook.

Shimizu, C., K.G. Nishimura, and T. Watanabe (2011), "House prices at different stages of the buying/selling process," Research Center for Price Dynamics Working Paper Series No. 69, February 2011.

Shimizu, C., K.G. Nishimura, and T. Watanabe (2010), "Housing prices in Tokyo: A comparison of hedonic and repeat-sales measures," *Journal of Economics and Statistics*, Volume 230, Issue 6, Special issue on "Index Theory and Price Statistics" edited by Erwin Diewert and Peter von der Lippe, December 2010, 792–813.