*JSPS Grants-in-Aid for Creative Scientific Research*
*Understanding Inflation Dynamics of the Japanese Economy*
*Working Paper Series No.69*

# House Prices at Different Stages of the Buying/Selling Process

Chihiro Shimizu
Kiyohiko G. Nishimura
and
Tsutomu Wanatabe

First draft: January 30, 2011
This version: April 28, 2011

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/

# House Prices at Different Stages of the Buying/Selling Process

Chihiro Shimizu* Kiyohiko G. Nishimura† Tsutomu Watanabe‡

First draft: January 30, 2011
This version: April 27, 2011

**Abstract**

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 of house prices. The choice of the dataset has been regarded as critically important from a practical viewpoint, but has not been discussed much in the literature. This study seeks to fill this gap by comparing the distributions 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 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. These four prices are collected by different parties and recorded in different datasets. We find 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, only small differences remain between the different house 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.

*JEL Classification Number*: R21; R31; C10
*Keywords*: house price index; quantile regressions; hedonic regressions; quality adjustment; goodness-of-fit tests

---

*Correspondence: Chihiro Shimizu, Reitaku University, Kashiwa, Chiba 277-8686, Japan. E-mail: cshimizu@reitaku-u.ac.jp. This paper is based on the outcome of a project at the Ministry of Land, Infrastructure, Transport and Tourism ("Study Group on the Development of Real Estate Price Indexes"), as well as on the outcome of a project at the Real Estate Information Network for East Japan ("Working Team on Research-Oriented Use of the REINS Database"). We would like to thank Erwin Diewert, David Fenwick, Sadao Sakamoto, and Hiwon Yoon for helpful discussions and comments. This research forms part of the project on "Understanding Inflation Dynamics of the Japanese Economy" funded by a JSPS Grant-in-Aid for Creative Scientific Research (18GS0101). Nishimura's contribution was made mostly before he joined the Policy Board of the Bank of Japan.

†Deputy Governor, Bank of Japan.

‡Hitotsubashi University and University of Tokyo.

# 1 Introduction

In constructing a housing price index, one has to make several nontrivial choices. One of them is the choice among alternative estimation methods, such as repeat-sales regression, hedonic regression, and so on. 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 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.[1] Needless to say, different datasets contain different types of prices, including sellers' asking prices, transactions prices, valuation prices, and so on.

With multiple datasets available, one may ask several questions. Are these prices different? If so, how do they differ from each other? 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

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

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.

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 indexes 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 finalized, 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.[2] 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 only use 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 similar distribution as 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 should not be affected by differences in house attributes between the two datasets. This idea is quite similar to the one adopted in the repeat sales method,

[2] An alternative approach would be to compare the four prices in terms of their average prices or in terms of their median prices. However, these statistics capture only one aspect of cross-sectional price distributions.

which is extensively used in constructing quality-adjusted house price indexes. 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.[3] 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.

The second method is based on hedonic regressions. This is again widely used in constructing quality-adjusted house price indexes. The hedonic regression we will employ in this paper differs from those extensively used in previous studies, which are based on the assumption that the hedonic coefficient on, say, the size of a house is identical for high-priced and low-priced houses. This restriction on hedonic coefficients may not be problematic as long as one is interested in the mean or the median of a price distribution, but it is a serious problem when one is interested in the shape of the entire price distribution. In this paper, we will use quantile hedonic regression in which hedonic coefficients are allowed to differ for high-priced and low-priced houses.

The main findings of this paper are as follows. We find that the four prices have substantially different distributions. However, these differences mainly come from differences in the attributes of houses contained in the different datasets. By looking at the intersections of the datasets and by employing quantile regressions, we show that once quality differences are eliminated, there remain only small differences between the price distributions.[4] These empirical results suggest 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.

[3]Shimizu et al. (2010) construct five different house price indexes, including hedonic and repeat sales indexes, using Japanese data for 1986 to 2008. They find that there exists a substantial discrepancy in terms of turning points between hedonic and repeat sales indexes. Specifically, the repeat sales measure signals turning points later than the hedonic measure: for example, the hedonic measure of condominium prices bottomed out at the beginning of 2002, while the corresponding repeat sales measure exhibits a reversal only in the spring of 2004.

[4]However, we find that the goodness-of-fit tests still reject the null that prices at different stages of the buying/selling process come from an identical distribution. Specifically, prices at earlier stages of the transaction process, such as asking prices initially listed in the magazine, tend to be slightly higher than prices at later stages, such as registry prices. This may reflect the fact that prices were updated downward in the transaction process due to weak demand in the Japanese housing market.

The rest of the paper is organized as follows. Section 2 outlines the data and the empirical methodology used in the paper. Section 3 provides the empirical results, and Section 4 concludes.

# 2 Data and Empirical Methodology

## 2.1 Data

In this paper, we focus on the prices of condominiums traded in the Greater Tokyo Area from September 2005 to December 2009.[5] 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 only 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 percent 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 withdraws 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 contract 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

[5]See Chapter 11 of Eurostat (2011) for detailed information on house price datasets currently available in Japan.

122,547, for a coverage of 34 percent. 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 is not 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 percent. 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 that these two prices 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. Summary statistics for the three datasets are presented in Table 1.

Some housing units appear only in 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.[6] 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 that the three datasets produce three different house price indexes, which behave quite differently, even if the identical estimation method is applied to each of the three datasets.

[6] The number of housing units that appear both in the realtor dataset and in the registry dataset is 22,613; the number of housing units that are in the realtor dataset but not in the registry dataset is 99,934; and the number of housing units that are in the registry dataset but not in the realtor dataset is 36,336.

## 2.2 The four prices at different stages of the house buying/selling process

Figure 1 shows the timing at which each of the four prices, $P_1$, $P_2$, $P_3$, and $P_4$, is observed in the house buying/selling process in Japan. There is a time lag of 70 days, on average, between the time when $P_1$ is observed (i.e., the time at which a seller posts an initial asking price in the magazine) and the time when $P_2$ is observed (i.e., the time when an offer is made by a buyer and the listing is removed from the magazine). Similarly, there is a lag of 38 days between the time at which $P_3$ is observed (i.e., the time at which a mortgage is approved and a contract is made) and the time at which $P_2$ is observed. Finally, there is a lag of 108 days between the time at which $P_4$ is observed (i.e., the time at which the MLIT receives price information from a buyer) and the time at which $P_3$ is observed.[7] In total, the time lag between $P_1$ and $P_4$ is, on average, 216 days, implying that a house price index can be available to the public much earlier by using $P_1$ instead of $P_4$. At the same time, it is often suggested that prices at the earlier stages of the house buying/selling process, such as $P_1$, are not reliable since they are frequently updated up or down until a final contract price is determined between the buyer and the seller. In addition, it is often pointed out that not all of the prices observed at the earlier stage of the house buying/selling process end in transactions.[8]

Figure 2 shows how time lags are distributed for the four prices. For example, the solid line represents the distribution of the time lag between the day $P_1$ and the day $P_2$ are observed for a particular property. We see that more than fifty percent of all observations are concentrated at a time lag of 50 days, but there is a non-negligible

[7] According to registry information, the day on which $P_3$ is observed and the day on which registration is made at the land registry are identical for 93 percent of all transactions. This means that the time lag between $P_3$ and $P_4$ mainly reflects the number of days it takes for the MLIT to collect price information from buyers. Note that this type of time lag does not occur in most other industrialized countries, including the U.S. and the U.K., where the land registry requires sellers and/or buyers to report transaction prices as part of the registration information. However, according to Eurostat (2011), even in the U.K., there exists a long time lag regarding the registration of property ownership transfers; that is, registration is typically completed only 4-6 weeks after the completion of transactions. This lack of timeliness means that price information gathered from registration is of limited usefulness in constructing house price indexes.

[8] Eurostat (2011: 147), for example, notes: "Each source of prices information has its advantages and disadvantages. For example a disadvantage of advertised prices and prices on mortgage applications and approvals is that not all of the prices included end in transactions, and the price may differ from the final negotiated transaction price. But these prices are likely to be available sometime before the final transaction price. Indices that measure the price earlier in the purchase process are able to detect price changes first, but will measure final prices with error because prices can be renegotiated extensively before the deal is finalized."

probability that the time lag exceeds 150 days. Similarly, the time lag between $P_1$ and $P_4$ is most likely to be 200 days, but it is possible, although with a low probability, that it may be more than 300 days.

### 2.3 Price distributions

Figure 3 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 4 shows the distributions of floor space, measured in square meters, for the three datasets. The distribution labeled "$P_1$ and $P_2$," which is from the magazine dataset, is almost symmetric, while the distribution labeled "$P_3$," 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 meters or less. This pattern is even more pronounced in the registry dataset, i.e., the distribution labeled "$P_4$." Turning to the middle and bottom panels of Figure 4, 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.

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 3 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.

### 2.4 Quality adjustment

We will adjust for quality in two different ways. For this purpose, let us begin by considering the dataset of initial asking prices in the magazine and the dataset of registry prices. We would like to conduct a quality adjustment before comparing the two price distributions. Let $F_1(p)$ denote the cumulative distribution function (CDF) of the log of initial asking prices in the magazine and $F_1(p \mid z)$ denote the conditional CDF of the log of initial asking prices, given a vector of house attributes, $z$. $F_1(p)$ and $F_1(p \mid z)$ are related as follows:

$$F_1(p) = \int_{-\infty}^{\infty} F_1(p \mid z) u_1(z) dz \tag{1}$$

where $u_1(z)$ is the distribution of $z$ for houses in the magazine dataset. Similarly, we define $F_4(p)$ and $F_4(p \mid z)$ for the log of registry prices, and $u_4(z)$ for houses in the registry dataset. Then we can express the difference between $F_1(p)$ and $F_4(p)$ as follows:

$$F_1(p) - F_4(p) = \int_{-\infty}^{\infty} \big[F_1(p \mid z) - F_4(p \mid z)\big] u_1(z) dz + \int_{-\infty}^{\infty} F_4(p \mid z) \big[u_1(z) - u_4(z)\big] dz \tag{2}$$

The conditional distribution of prices given $z$ can be interpreted as the distribution of quality adjusted prices, so that the first term on the right hand side of eqn. (2) represents the difference between the two distributions of quality adjusted prices. According to eqn. (2), however, the difference between $F_1(p)$ and $F_4(p)$ also comes from differences in terms of the distribution of $z$ between the two datasets (i.e., the magazine dataset and the registry dataset), which is represented by the second term on the right hand side. Our aim is to eliminate the second term before comparing the two price distributions.

**Intersection approach** The first method to do so is to use prices only for houses that are in both the magazine and the registry datasets. We will refer to this as the "intersection approach." We use address information to identify these houses. The number of houses in this intersection sample turns out to be 15,015. Since the distributions $u_1(z)$ and $u_4(z)$ are identical in the intersection sample, eqn. (2) becomes

$$F_1(p) - F_4(p) = \int_{-\infty}^{\infty} \big[F_1(p \mid z) - F_4(p \mid z)\big] u_1(z) dz. \tag{3}$$

This is how we eliminate the effect of quality differences between datasets.

Note that this method is based on the same idea as the repeat sales method, which is extensively used in constructing quality adjusted house price indexes. As is often pointed out, however, repeat sales samples may not necessarily be representative samples, because houses that are traded multiple times may have some special characteristics that set them apart from other houses. A similar type of sample selection bias may arise in our intersection approach. Houses in the intersection between the magazine dataset and the registry dataset are cases which successfully ended in a transaction. In other words, 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. If such unsuccessful transactions occur randomly, then this would not pose a problem. However, if unsuccessful transactions are more frequent for, say, high-priced houses, this would give rise to sample selection bias, since the distribution of initial asking prices and that of registry prices will differ.

**Quantile hedonic approach** The second method for quality adjustment is based on quantile hedonic regression. Let $Q_i^\theta(p \mid z)$ denote the $\theta$-th quantile of the conditional distribution $F_i(p \mid z)$, where $\theta \in (0, 1)$. Following Machado and Mata (2005), we model these conditional quantiles by

$$Q_i^\theta(p \mid z) = z\beta_i(\theta) \tag{4}$$

This simply states that the conditional quantiles are a weighted average of various house attributes. This is similar to the idea of hedonic regressions, but differs from them in that the weight vector, $\beta_i(\theta)$, is assumed to depend on the value of $\theta$. The restriction that hedonic coefficients do not depend on quantiles may not be problematic as long as one is interested in the mean or in the median of the distribution of quality adjusted prices, but it is a serious problem when one is interested in the shape of the entire distribution of quality adjusted prices. We will eliminate this restriction by employing quantile hedonic regression. Note that the weight vector $\beta_i(\theta)$, which is referred to as the quantile regression coefficient, can be interpreted as the shadow price associated with each of the house attributes.

Given this setting, we proceed as follows. We first apply quantile regression to initial asking prices and house attributes in the magazine dataset to obtain the estimate of $\beta_1(\theta)$, which is denoted by $\hat{\beta}_1(\theta)$. Given the values of $z$ and $p$, we are then able to calculate $F_1(p \mid z)$ using the equation $p = z\hat{\beta}_1(\theta)$. The estimate of $F_1(p \mid z)$ is denoted by $\hat{F}_1(p \mid z)$. Similarly, we obtain the estimate of $F_4(p \mid z)$, denoted by $\hat{F}_4(p \mid z)$, using the registry dataset. By integrating out $z$, we obtain the estimates of the marginal distributions as follows:

$$\hat{F}_1(p) \equiv \int_{-\infty}^{\infty} \hat{F}_1(p \mid z)u_1(z)dz; \quad \hat{F}_4(p) \equiv \int_{-\infty}^{\infty} \hat{F}_4(p \mid z)u_4(z)dz \tag{5}$$

Then we have an equation analogous to eqn. (2):

$$\hat{F}_1(p) - \hat{F}_4(p) = \int_{-\infty}^{\infty} \big[\hat{F}_1(p \mid z) - \hat{F}_4(p \mid z)\big] u_1(z) dz + \int_{-\infty}^{\infty} \hat{F}_4(p \mid z)\big[u_1(z) - u_4(z)\big] dz \tag{6}$$

We use the second term on the right hand side as the difference between the two distributions of quality adjusted prices.

Specifically, we estimate the distribution of quality adjusted prices closely following the method developed by Machado and Mata (2005), consisting of the following steps:

1. Estimate quantile regressions for $Q$ values of $\theta$. The estimates are $\hat{\beta}_1(\theta)$ for the magazine dataset and $\hat{\beta}_4(\theta)$ for the registry dataset.

2. Draw with replacement from the $Q$ sets of quantile regression coefficient vectors. The individual draws are denoted by $\hat{\beta}_1(b)$ and $\hat{\beta}_4(b)$, where $b = 1, \ldots, B$. A uniform distribution is used, i.e., each $\theta$ is equally likely to be drawn.

3. Draw with replacement from $z_{1j}$ and $z_{4k}$, where $z_{1j}$ is the vector of explanatory variables for observation $j$ in the magazine dataset ($j = 1, \ldots, n_1$), and $z_{4k}$ is the vector of explanatory variables for observation $k$ in the registry dataset ($k = 1, \ldots, n_4$). Each observation is equally likely to appear in the new vectors, $z_{1b}$ and $z_{4b}$, $b = 1, \ldots, B$.

4. Calculate $z_{1b}\hat{\beta}_1(b)$, $z_{4b}\hat{\beta}_4(b)$, and $z_{1b}\hat{\beta}_4(b)$.

5. Estimate the density functions for $z_{1b}\hat{\beta}_1(b)$, $z_{4b}\hat{\beta}_4(b)$, and $z_{1b}\hat{\beta}_4(b)$. The estimates correspond to $\hat{F}_1(p)$, $\hat{F}_4(p)$, and $\int_{-\infty}^{\infty} \hat{F}_4(p \mid z) u_1(z) dz$.

Machado and Mata (2005) use this method to decompose the change of the wage distribution over time into several factors, while McMillen (2008) was the first to apply this method to a cross-sectional house price distribution. Specifically, he decomposed the change in the house price distribution in Chicago between 1995 and 2005 into two components: the change in the distribution of house attributes and the change in the coefficients on the house attributes (i.e., the shadow prices associated with the attributes). His main finding is that the change in the price distribution over time comes mainly from the change in the coefficients on house attributes rather than the change in the distribution of house attributes. These two studies employ this method to investigate the difference between two distributions (i.e., wage distributions or house price distributions) at time $t$ and at time $t + 1$. The present study differs from these two papers in that we compare two distributions that come from different datasets rather than two distributions at different points in time from the same dataset.

# 3 Empirical Results

## 3.1 The distribution of quality adjusted prices: Results based on the intersection approach

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 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.

We start by looking at the distribution of relative prices between $P_1$ and $P_4$ and between $P_2$ and $P_4$ for housing units in the intersection of the magazine and registry datasets. Similarly, we look at the distribution of relative prices between $P_3$ and $P_4$ for housing units in the intersection of the realtor and registry datasets. Figure 5 shows that the distribution of $P_1/P_4$ has the largest density at a range of 1.05 to 1.10, with more than thirty percent of the total observations being concentrated in this range, and that the densities above 1.10 are not negligible. In contrast, the number of houses for which $P_4$ exceeds $P_1$ is very limited, indicating that initial asking prices tend to be higher than registry prices. This may reflect the weak housing demand in the period from 2005 to 2009 when the price data was collected. Turning to the relative price $P_2/P_4$, the densities for the range of 1.00 and 1.05, and the range of 1.05 and 1.10, are slightly higher than those for the relative price $P_1/P_4$, indicating that final asking prices listed in the magazine tend to be closer to registry prices than initial asking prices. This tendency is more clearly seen for the relative prices between realtor prices and registry prices: more than 70 percent of observations are concentrated in the range of 1.00 to 1.05 for $P_3/P_4$.

Next, Figure 6 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 3, 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 6 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 3, 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 7, 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 45 degree reference line. The greater the departure from this reference line, the more this suggests that the two sets of prices come from populations with different distributions.

The panels in Figure 7(a) show the q-q plots for raw prices, the distributions of which were shown in Figure 3. 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 $P_1$. 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 7(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.

To conduct a formal test to determine if the two distributions come from populations with the same distribution, we calculate the $D$ statistic, which measures the deviation between the two distributions and is defined as follows:

$$D = \max_c \mid F_x(c) - F_y(c) \mid \tag{7}$$

where $F_x(\cdot)$ and $F_y(\cdot)$ are cumulative distributions for two random variables. The estimated $D$'s are shown in Table 2. The results for the raw data are presented in the first three rows, while the results for the quality adjusted data using the intersection approach are shown in the next three rows. For example, concerning the distributions of $P_1$ and $P_4$, the estimate of $D$ is 0.2016 for the raw data, indicating that the two cumulative distributions deviate a substantial 20 percent. On the other hand, the estimate of $D$ for the quality adjusted data employing the intersection approach is 0.0584, much smaller than the corresponding value for the raw data. However, this does not necessarily mean that the two distributions are the same. In fact, we find that the $p$-value obtained from the Kolmogorov-Smirnov test (KS test) is very close to zero, implying that the null hypothesis that the two distributions are identical can be easily rejected not only for the raw data but also for the quality adjusted data. We also find that $D = 0.0441$ for the distributions of $P_2$ and $P_4$, and $D = 0.0303$ for the distributions of $P_3$ and $P_4$, implying that the deviation from the distribution of registry prices, $P_4$, becomes smaller and smaller at later (i.e., downstream) stages of the house buying/selling process, although the null hypothesis is still rejected for these cases.

One may wonder how the deviations between the four prices fluctuate over time. In particular, an important question to be asked is whether the deviations differ depending on whether the housing market is in a downturn or in an upturn. To address this, we present in Figure 8 a time series for the price ratio between $P_1$ and $P_2$, as well as a time series for the interval between the time when $P_1$ is observed (i.e., the time at which a seller posts an initial asking price in the magazine) and the time when $P_2$ is observed (i.e., the time when an offer is made by a buyer and the listing is removed from the magazine).

The price ratio for a particular month is defined and calculated as the average of the ratios between $P_2$ and $P_1$ for housing units for which an offer is made in that month and for which an initial asking price $P_1$ was listed in the magazine some time prior to that month. As shown in the lower panel of Figure 8, the price ratio fluctuates between 0.97 and 0.99, indicating that $P_2$ tends to be lower than $P_1$ by one to three percent. More importantly, we see that fluctuations in the price ratio are closely correlated with the overall price movement in the housing market, which is represented by the hedonic indexes for $P_1$ and $P_2$ shown in the upper panel of Figure 8. Specifically, the hedonic index for $P_1$ declined by more than ten percent during the period between March 2008 and April 2009 indicated by the shaded area. During this downturn period, the price ratio exhibited a substantial decline, and more interestingly, changes in the price ratio preceded changes in the hedonic indexes. Specifically, the price ratio started to decline in December 2007, three months earlier than the hedonic index for $P_1$, and bottomed out in February 2009, two months earlier than the hedonic index for $P_1$.

Next, the interval for a particular month is calculated as the average of the time lags between the time $P_2$ is observed and the time $P_1$ is observed for those housing units for which an offer is made in that month. The interval fluctuates between 55 and 78 days, and more importantly, it is closely correlated with the hedonic indexes for $P_1$ and $P_2$. Focusing on the downturn period, which is indicated by the shaded area, we see that the interval increased from 65 days to 78 days, suggesting that, due to weak demand, sellers had to wait longer until an offer is made by a buyer. As in the case of the price ratio, changes in the interval tended to precede changes in the hedonic indexes; specifically, the interval peaked in December 2008, four months before than the hedonic index for $P_1$ hit bottom.

### 3.2 The distribution of quality adjusted prices: Results based on the quantile hedonic approach

We first conduct a standard hedonic regression for (the log of) the four prices using a similar specification as the one adopted by Shimizu et al. (2010) and others. A list of the variables used in the regression is provided in Table 3. The regression results are shown in the first two columns of Table 4. The results are standard: house prices increase with floor space and decline with age, distance to the nearest station, and commuting time to the central business district; in addition, prices are higher for houses with main windows facing south, and for houses with a steel reinforced concrete frame structure. There are some differences across the four prices in the estimated coefficients, but they are not very large. Each of the four regressions explains more than 70 percent of the variation in the log of house price.

To conduct the Machado-Mata (2005) decomposition, we run 97 quantile regressions for quantiles ranging from $\theta = 0.02$ to 0.98 in increments of 0.01. Table 4 shows the estimates for representative (25 percent, 50 percent, and 75 percent) quantiles, while Figure 9 shows the regression coefficients by quantile. We see that several variables exhibit significant quantile effects. Specifically, the estimated coefficient on the age of a building is negative but tends to become closer to zero at higher quantiles, implying that age decreases house prices, but less so for high-priced houses. The coefficient on the distance to the nearest station is close to zero for quantiles above 90 percent, indicating that the distance to the nearest station is a less important determinant of price for high-priced houses. There is a similar tendency for the coefficient on commuting time to the central business district. On the other hand, we see no significant quantile effects for floor space; that is, the coefficient on floor space does not change much at different quantiles.

Comparing the estimated coefficients for the four prices, we do not see any significant differences. For example, the coefficient on age differs between the four prices in that it is closer to zero for $P_1$ and $P_2$ than for the other two prices, but the difference is not statistically significant. Similarly, the coefficient on floor space is smaller (closer to zero) for $P_1$ and $P_2$ than for the other two prices, but the difference is not very large.

Based on the results from the quantile hedonic regressions, we decompose the difference between price distributions into two components: the difference resulting from differences in quantile regression coefficients, and the difference resulting from differences in house attribute distributions. For each of the four prices, we make 50,000 independent draws from the vectors of house attributes and 50,000 independent draws from the estimated quantile coefficient vectors. We then use the results to estimate the density functions for $z_{ib}\hat{\beta}_i(b)$ and $z_{ib}\hat{\beta}_4(b)$ for $i = 1, 2, 3, 4$.

The results of this exercise are presented in Figure 10. The top, middle, and bottom panels respectively show the results for the difference between $P_1$ and $P_4$, between $P_2$ and $P_4$, and between $P_3$ and $P_4$. The solid lines show the difference between the two price distributions, while the dashed lines show the difference due to differences in the quantile hedonic coefficients and the difference due to differences in the distribution of house attributes. For example, the solid line in the upper panel represents $\hat{f}_1(p) - \hat{f}_4(p)$, where $\hat{f}_i(p)$ is the estimated density function whose CDF is given by $\hat{F}_i(p)$. The dashed line labeled "coefficients" in the top panel represents the contribution of differences in the quantile hedonic coefficients, which is given by

$$\int_{-\infty}^{\infty} \left[\hat{f}_1(p \mid z) - \hat{f}_4(p \mid z)\right] u_1(z) dz$$

where $\hat{f}_i(p \mid z)$ is the estimated conditional density function whose CDF is given by $\hat{F}_i(p \mid z)$, while the dashed line labeled "variables" in the top panel represents the contribution of differences in the distribution of house attributes, which is given by

$$\int_{-\infty}^{\infty} \hat{f}_4(p \mid z) \left[u_1(z) - u_4(z)\right] dz.$$

The solid and dashed lines in the other two panels are similarly defined.

The top panel of Figure 10 shows that the difference between the distributions of (the log of) $P_1$ and $P_4$ mainly comes from differences in the distribution of house attributes. However, differences in the quantile hedonic coefficient also make a certain contribution. In other words, there remain non-trivial differences between the price distributions even after prices are quality adjusted using the quantile hedonic approach. In the middle and bottom panels, the contribution of differences in the quantile hedonic coefficients becomes smaller but still seems to be non-negligible.[9]

Next, we again use q-q plots to see how close the distributions are. The top panel in Figure 7(c) compares the distribution of quality adjusted values of (the log of) $P_1$, which are defined by $\int_{-\infty}^{\infty} \hat{f}_1(p \mid z)u_1(z)dz$, and the distribution of quality adjusted values of (the log of) $P_4$, which are defined by $\int_{-\infty}^{\infty} \hat{f}_4(p \mid z)u_1(z)dz$. The other two panels in Figure 7(c) compare the distribution of quality adjusted values for $P_2$ and $P_4$ and for $P_3$ and $P_4$. We see in the top and middle panels that the dots are located above the 45 degree line, especially at lower quantiles, indicating that there remains some difference between the price distributions even after quality differences are controlled for by quantile hedonic regression. On the other hand, we see little deviation from the 45 degree line in the bottom panel showing the q-q plot for $P_3$ and $P_4$.

Finally, we conduct KS tests to determine if the price distributions come from populations with a common distribution. The results are presented in the bottom three rows of Table 2. The estimated values of the $D$ statistic are $D = 0.0676$ for the distributions of $P_1$ and $P_4$, $D = 0.0535$ for the distributions of $P_2$ and $P_4$, and $D = 0.0199$ for the distributions of $P_3$ and $P_4$, suggesting that the distance between the price distributions tends to be smaller when two sets of prices come from closer stages of the house buying/selling process. However, the $p$-values associated with the three KS tests are all very close to zero, implying that the null hypothesis that the price distributions come from populations with a common distribution is easily rejected even after prices are quality adjusted using the hedonic quantile approach.[10]

[9]These results are quite different from the ones reported by McMillen (2008), who compared two price distributions, one for 1995 and the other for 2005, and found that virtually the entire difference between the price distributions was due to differences in the quantile hedonic coefficients. Note, however, that McMillen (2008) compared price distributions from different years, while we compare price distributions from different stages of the house buying/selling process, and therefore from different datasets compiled by different parties.

[10]Note that the quality adjustment using the intersection approach tends to produce lower estimates of the $D$ statistic than the quality adjustment using the quantile hedonic approach. For example, the estimate of $D$ for $P_1$ vs. $P_4$ is 0.0584 for the intersection approach but 0.0676 for the quantile hedonic approach. As discussed earlier, quality adjustment using the intersection approach may suffer from sample selection bias because of unsuccessful transactions (i.e., some houses are listed in the magazine, but not recorded in the registry dataset, for example, because the potential buyer was

## 4 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 of 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 by employing quantile hedonic regressions as proposed by Machado and Mata (2005), 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.

[To be completed]

## References

[1] Eurostat (2011), *Handbook on Residential Property Price Indices*, Draft Version 3.0, January 2011. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/hicp/methodology/owner_occupied_housing_hpi/rppi_handbook.

[2] McMillen, D. P. (2008), "Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients?" *Journal of Urban Economics* 64, 2008, 573-589.

[3] Koenker, R. (2005), *Quantile Regression.* Cambridge Univ. Press, New York.

unable to obtain a mortgage). In other words, the intersection samples contain prices only from very homogeneous (maybe excessively homogeneous) houses.

[4] Koenker, R. and K. F. Hallock (2001), "Quantile regression," *Journal of Economic Perspectives* 15, 2001, 143-156.

[5] Machado, J. A. F. and J. Mata (2005), "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics* 20, 445-465.

[6] Ohnishi, T., T. Mizuno, C. Shimizu, and T. Watanabe (2010), "On the evolution of the house price distribution," Research Center for Price Dynamics Working Paper Series No. 61, August 2010.

[7] Pollakowski, H. O. (1995), "Data sources for measuring house price changes," Journal of Housing Research 6(3), 377-387.

[8] 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.

# Table 1: Summary of the three datasets

**Magazine data ($P_1$, $P_2$)**

| | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| | Magazine data (155,347 observations) | | | |
| $P_1$: First asking price (10,000 Yen) | 2,958.51 | 1,875.16 | 200 | 33,000 |
| $P_2$: Final asking price (10,000 Yen) | 2,889.27 | 1,831.34 | 200 | 29,800 |
| Log $P_1$: Log of $P_1$ | 7.84 | 0.54 | 5.77 | 10.40 |
| Log $P_2$: Log of $P_2$ | 7.82 | 0.54 | 5.77 | 10.30 |
| $FS$: Floor space (m$^2$) | 66.77 | 18.97 | 10.39 | 243.90 |
| $P_1$ / $FS$ (10,000 Yen) | 43.59 | 21.65 | 10.87 | 195.68 |
| $P_2$ / $FS$ (10,000 Yen) | 42.58 | 21.16 | 10.00 | 189.08 |
| $AGE$: Age of building (years) | 16.59 | 10.26 | 1.50 | 58.93 |
| $DS$: Distance to the nearest station (meters) | 850.42 | 729.86 | 80 | 9,900 |
| $TT$: Travel time to terminal station (minutes) | 20.97 | 12.61 | 2 | 89 |

**Realtor data ($P_3$)**

| | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| | Realtor data (122,547 observations) | | | |
| $P_3$: Sales price (10,000 Yen) | 2,431.81 | 1,632.88 | 160 | 29,074 |
| Log $P_3$: Log of $P_3$ | 7.60 | 0.64 | 5.08 | 10.28 |
| $FS$: Floor space (m$^2$) | 64.87 | 20.27 | 10.10 | 238.81 |
| $P_3$ / $FS$ (10,000 Yen) | 37.44 | 19.65 | 10.00 | 187.96 |
| $AGE$: Age of building (years) | 16.79 | 10.38 | 1.50 | 57.14 |
| $DS$: Distance to the nearest station (meters) | 881.37 | 804.67 | 80 | 9,900 |
| $TT$: Travel time to terminal station (minutes) | 23.21 | 13.65 | 2 | 89 |

**Registry data ($P_4$)**

| | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|
| | Registry data (58,949 observations) | | | |
| $P_4$: Sales price (10,000 Yen) | 2,316.32 | 1,633.34 | 130 | 28,000 |
| Log $P_4$: Log of $P_4$ | 7.53 | 0.68 | 4.87 | 10.24 |
| $FS$: Floor space (m$^2$) | 57.52 | 23.58 | 10.09 | 196.46 |
| $P_4$ / $FS$ (10,000 Yen) | 41.38 | 21.53 | 10.00 | 189.83 |
| $AGE$: Age of building (years) | 16.21 | 9.83 | 1.50 | 59.40 |
| $DS$: Distance to the nearest station (meters) | 842.77 | 719.73 | 50 | 9,910 |
| $TT$: Travel time to terminal station (minutes) | 21.23 | 13.51 | 2 | 89 |

## Table 2: Goodness-of-fit tests

| | $D$-statistic | $p$-value | Number of observations |
|---|---|---|---|
| | | *Raw data* | |
| $P_1$ vs. $P_4$ | 0.2016 | 0.000 | 155,347 for $P_1$ and 58,949 for $P_4$ |
| $P_2$ vs. $P_4$ | 0.1885 | 0.000 | 155,347 for $P_2$ and 58,949 for $P_4$ |
| $P_3$ vs. $P_4$ | 0.0432 | 0.000 | 122,547 for $P_3$ and 58,949 for $P_4$ |
| | | *Quality adjusted by the intersection approach* | |
| $P_1$ vs. $P_4$ | 0.0584 | 0.000 | 14,890 for $P_1$ and 14,890 for $P_4$ |
| $P_2$ vs. $P_4$ | 0.0441 | 0.000 | 14,890 for $P_2$ and 14,890 for $P_4$ |
| $P_3$ vs. $P_4$ | 0.0303 | 0.000 | 22,613 for $P_3$ and 22,613 for $P_4$ |
| | | *Quality adjusted by the quantile hedonic approach* | |
| $P_1$ vs. $P_4$ | 0.0676 | 0.000 | 50,000 for $P_1$ and 50,000 for $P_4$ |
| $P_2$ vs. $P_4$ | 0.0535 | 0.000 | 50,000 for $P_2$ and 50,000 for $P_4$ |
| $P_3$ vs. $P_4$ | 0.0199 | 0.000 | 50,000 for $P_3$ and 50,000 for $P_4$ |

# Table 3: List of variables

| Symbol | Variable | Content | Unit |
|---|---|---|---|
| *FS* | Floor space | Floor space of building/square meters | $m^2$ |
| *AGE* | Age of building at the time of transaction | Age of building at the time of transaction. | years |
| *DS* | Distance to the nearest station | Distance to the nearest station. | meters |
| *TT* | Travel time to terminal station | Minimum railway riding time in daytime to one of the seven major business district stations. | minutes |
| *SD* | South-facing dummy | Main windows facing south = 1 | (0,1) |
| | | Main windows not facing south = 0 | |
| *SRC* | Steel reinforced concrete dummy | Steel reinforced concrete frame structure = 1 | (0,1) |
| | | Other structure = 0 | |
| *STD* | Studio type dummy | Floor space 30 square meters or less =1 | (0,1) |
| | | Floor space over 30 square meters = 0 | |
| $LD_k$ $(k=0,\dots,K)$ | Location (ward) dummy | $k$-th administrative district =1, | (0,1) |
| | | Other district =0. | |
| $RD_l$ $(l=0,\dots,L)$ | Railway line dummy | $l$-th railway line =1 | (0,1) |
| | | Other railway line = 0. | |
| $TD_m$ $(m=0,\dots,M)$ | Time dummy (yearly) | $m$-th year =1 | (0,1) |
| | | Other year =0. | |

# Table 4: Quantile regressions

| $P_1$ (155,347 observations) | OLS Coef. | OLS Std. err. | 25% Coef. | 25% Std. err. | 50% Coef. | 50% Std. err. | 75% Coef. | 75% Std. err. |
|---|---|---|---|---|---|---|---|---|
| Constant | 7.445 | 0.004 | 7.261 | 0.005 | 7.330 | 0.005 | 7.528 | 0.006 |
| *FS* : Floor space ($m^2$) | 0.015 | 0.000 | 0.015 | 0.000 | 0.016 | 0.000 | 0.016 | 0.000 |
| *AGE* : Age of building (years) | -0.019 | 0.000 | -0.021 | 0.000 | -0.019 | 0.000 | -0.017 | 0.000 |
| *DS* : Distance to the nearest station (meters) | -0.00015 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *TT* : Travel time to terminal station (minutes) | -0.017 | 0.000 | -0.016 | 0.000 | -0.016 | 0.000 | -0.018 | 0.000 |
| *SD* : South dummy | 0.013 | 0.001 | 0.012 | 0.002 | 0.013 | 0.002 | 0.011 | 0.002 |
| *SRC* : Steel reinforced concrete dummy | 0.017 | 0.001 | 0.030 | 0.002 | 0.018 | 0.002 | 0.005 | 0.002 |
| *STD* : Studio type dummy | -0.412 | 0.005 | -0.436 | 0.007 | -0.376 | 0.006 | -0.334 | 0.006 |
| Adjusted R-squared= | 0.784 | | | | | | | |

| $P_2$ (155,347 observations) | OLS Coef. | OLS Std. err. | 25% Coef. | 25% Std. err. | 50% Coef. | 50% Std. err. | 75% Coef. | 75% Std. err. |
|---|---|---|---|---|---|---|---|---|
| Constant | 7.434 | 0.004 | 7.253 | 0.005 | 7.323 | 0.006 | 7.518 | 0.006 |
| *FS* : Floor space ($m^2$) | 0.015 | 0.000 | 0.015 | 0.000 | 0.016 | 0.000 | 0.016 | 0.000 |
| *AGE* : Age of building (years) | -0.019 | 0.000 | -0.022 | 0.000 | -0.019 | 0.000 | -0.017 | 0.000 |
| *DS* : Distance to the nearest station (meters) | -0.00015 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *TT* : Travel time to terminal station (minutes) | -0.018 | 0.000 | -0.017 | 0.000 | -0.016 | 0.000 | -0.018 | 0.000 |
| *SD* : South dummy | 0.014 | 0.001 | 0.013 | 0.002 | 0.012 | 0.002 | 0.011 | 0.002 |
| *SRC* : Steel reinforced concrete dummy | 0.019 | 0.002 | 0.031 | 0.002 | 0.020 | 0.002 | 0.005 | 0.002 |
| *STD* : Studio type dummy | -0.415 | 0.005 | -0.439 | 0.007 | -0.383 | 0.006 | -0.337 | 0.006 |
| Adjusted R-squared= | 0.784 | | | | | | | |

| $P_3$ (122,547 observations) | OLS Coef. | OLS Std. err. | 25% Coef. | 25% Std. err. | 50% Coef. | 50% Std. err. | 75% Coef. | 75% Std. err. |
|---|---|---|---|---|---|---|---|---|
| Constant | 7.327 | 0.005 | 7.148 | 0.006 | 7.243 | 0.006 | 7.438 | 0.008 |
| *FS* : Floor space ($m^2$) | 0.016 | 0.000 | 0.017 | 0.000 | 0.017 | 0.000 | 0.017 | 0.000 |
| *AGE* : Age of building (years) | -0.024 | 0.000 | -0.026 | 0.000 | -0.024 | 0.000 | -0.021 | 0.000 |
| *DS* : Distance to the nearest station (meters) | -0.00014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *TT* : Travel time to terminal station (minutes) | -0.017 | 0.000 | -0.016 | 0.000 | -0.016 | 0.000 | -0.018 | 0.000 |
| *SD* : South dummy | 0.023 | 0.002 | 0.021 | 0.002 | 0.016 | 0.002 | 0.022 | 0.002 |
| *SRC* : Steel reinforced concrete dummy | 0.032 | 0.002 | 0.045 | 0.002 | 0.032 | 0.002 | 0.014 | 0.002 |
| *STD* : Studio type dummy | -0.477 | 0.004 | -0.485 | 0.005 | -0.444 | 0.005 | -0.422 | 0.005 |
| Adjusted R-squared= | 0.830 | | | | | | | |

| $P_4$ (58,949 observations) | OLS Coef. | OLS Std. err. | 25% Coef. | 25% Std. err. | 50% Coef. | 50% Std. err. | 75% Coef. | 75% Std. err. |
|---|---|---|---|---|---|---|---|---|
| Constant | 7.242 | 0.009 | 7.092 | 0.009 | 7.189 | 0.010 | 7.357 | 0.013 |
| *FS* : Floor space ($m^2$) | 0.017 | 0.000 | 0.017 | 0.000 | 0.018 | 0.000 | 0.018 | 0.000 |
| *AGE* : Age of building (years) | -0.023 | 0.000 | -0.026 | 0.000 | -0.023 | 0.000 | -0.021 | 0.000 |
| *DS* : Distance to the nearest station (meters) | -0.00014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *TT* : Travel time to terminal station (minutes) | -0.015 | 0.000 | -0.015 | 0.000 | -0.015 | 0.000 | -0.015 | 0.000 |
| *SD* : South dummy | 0.016 | 0.003 | 0.008 | 0.003 | 0.012 | 0.003 | 0.015 | 0.004 |
| *SRC* : Steel reinforced concrete dummy | 0.031 | 0.003 | 0.041 | 0.003 | 0.037 | 0.003 | 0.024 | 0.003 |
| *STD* : Studio type dummy | -0.439 | 0.005 | -0.447 | 0.006 | -0.407 | 0.006 | -0.362 | 0.007 |
| Adjusted R-square= | 0.796 | | | | | | | |

Note: The dependent variable in each case is the log price.

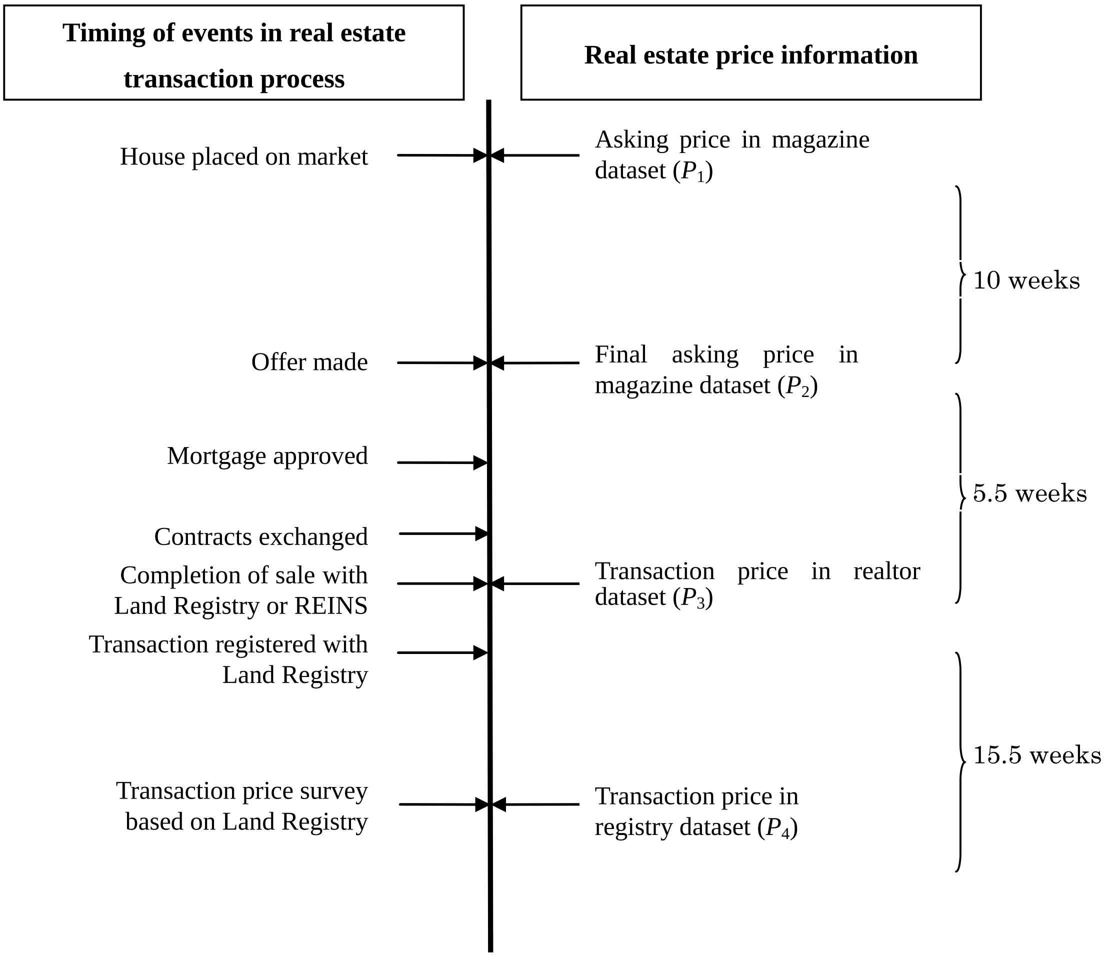



**Figure 1: House purchase timeline**

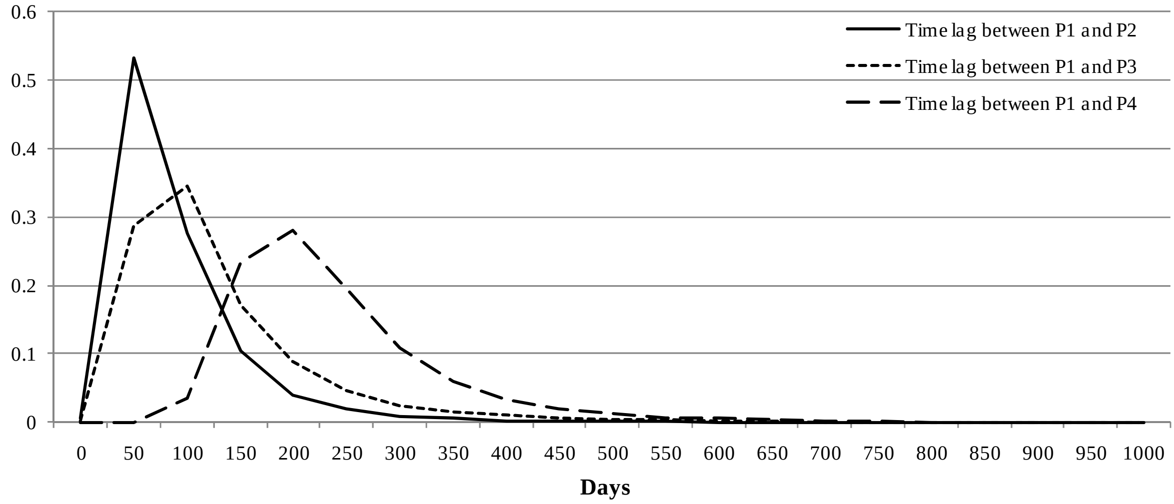



**Figure 2: Intervals between events in the house buying/selling process**

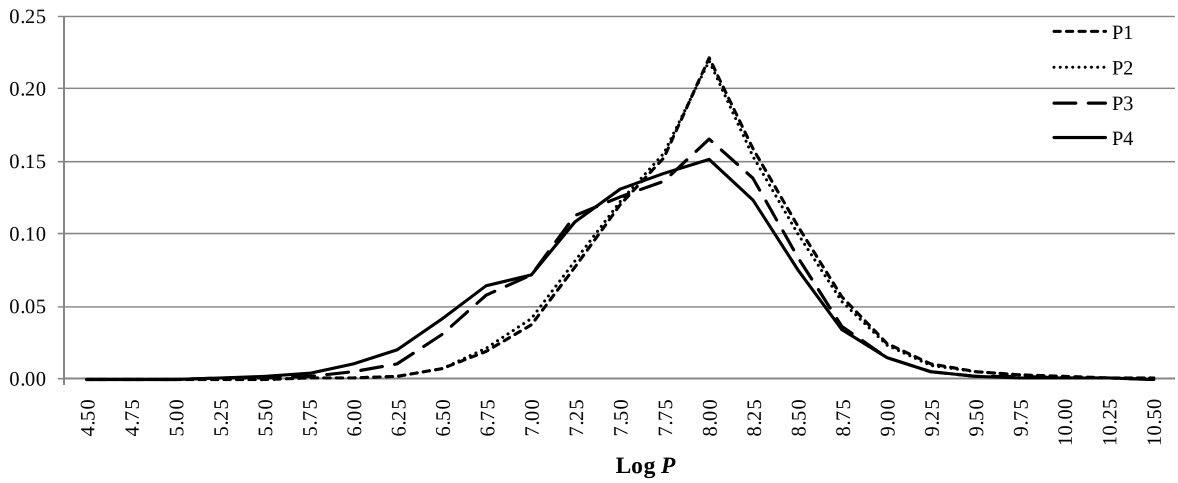
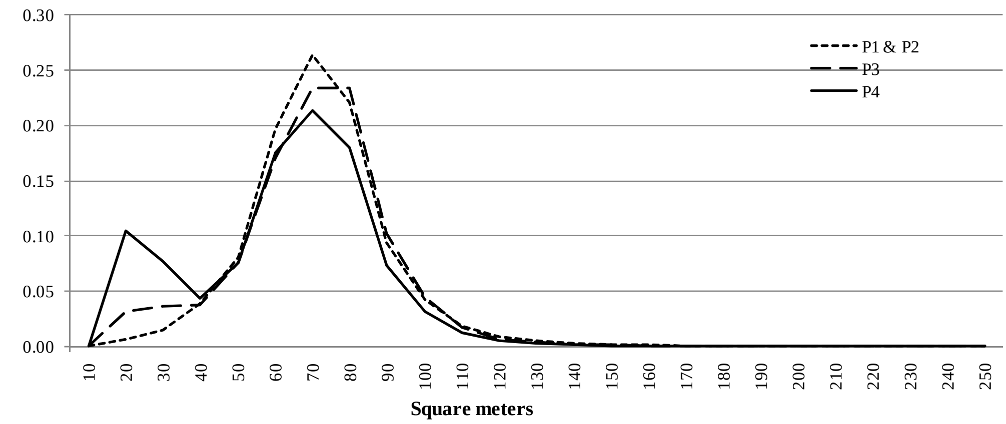
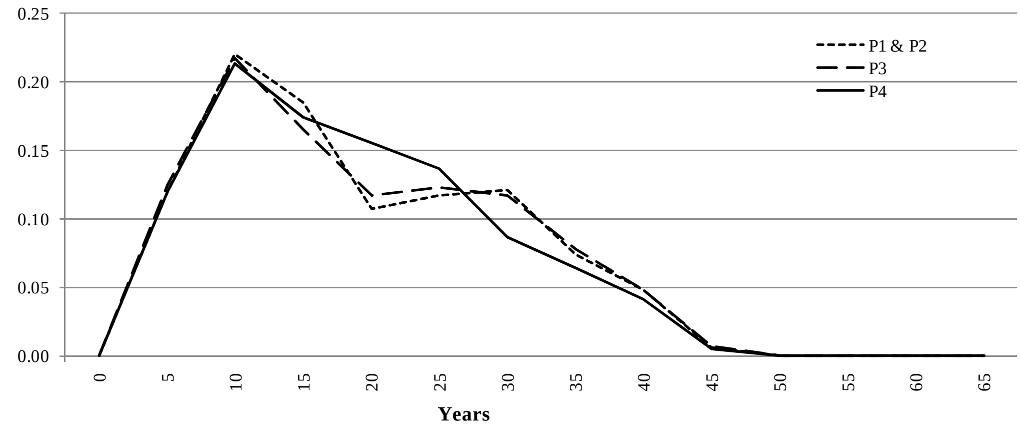


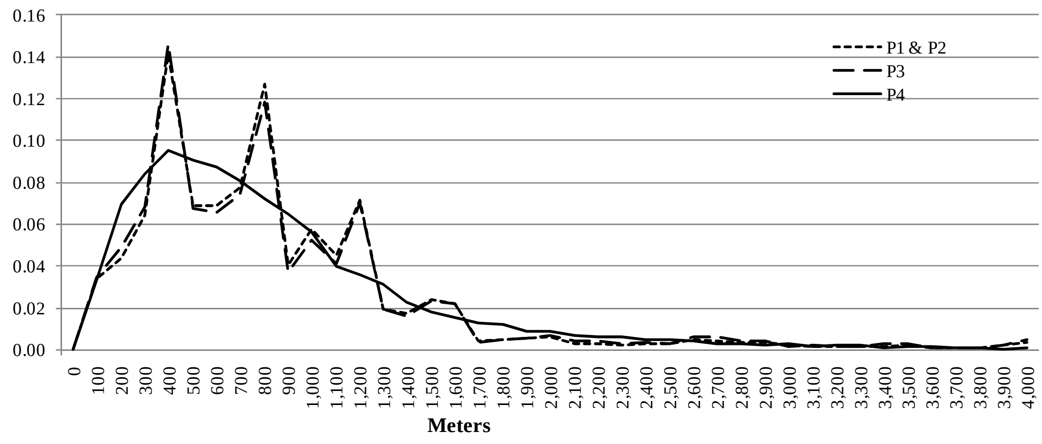

**Figure 3: Price densities for $P_1$, $P_2$, $P_3$, and $P_4$**




*FS* : Floor space




*AGE* : Age of building




*DS* : Distance to the nearest station

**Figure 4: Density functions for house attributes**

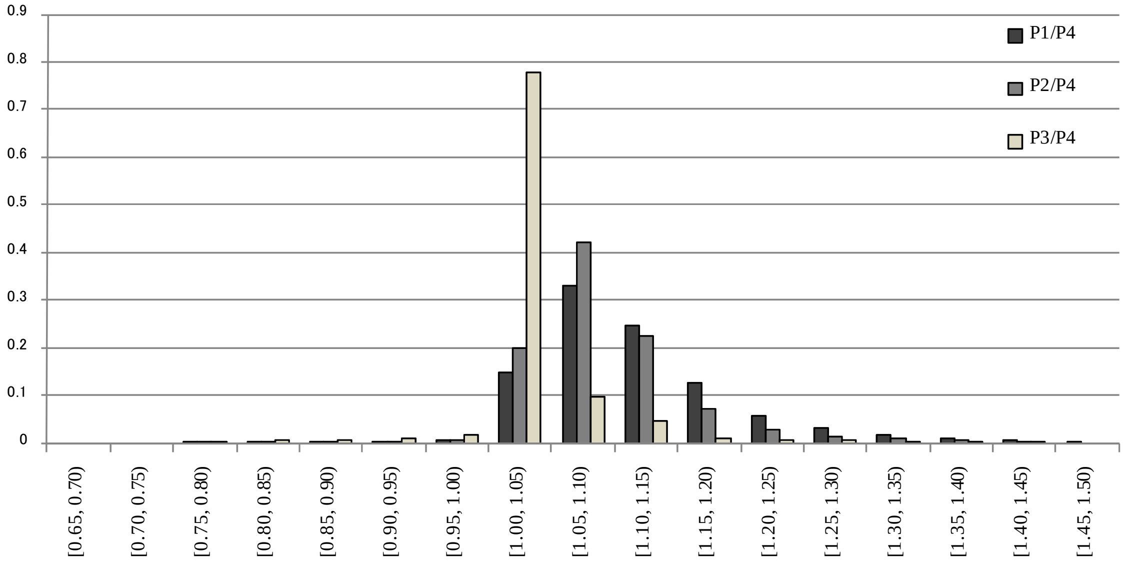
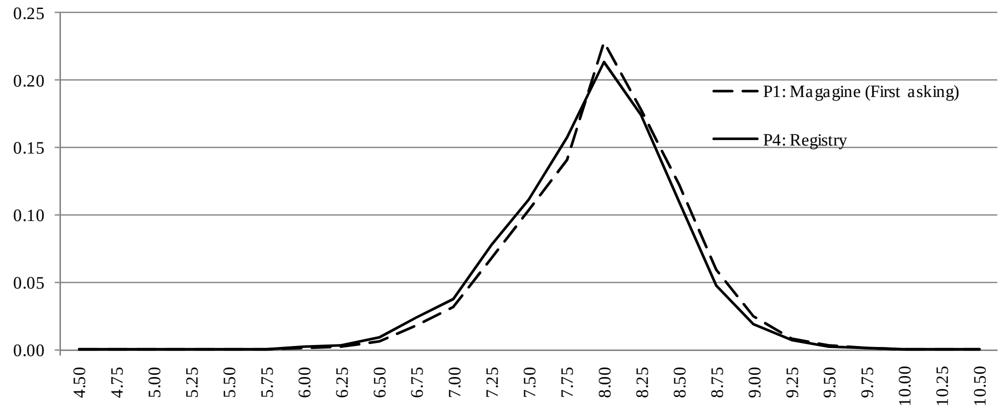
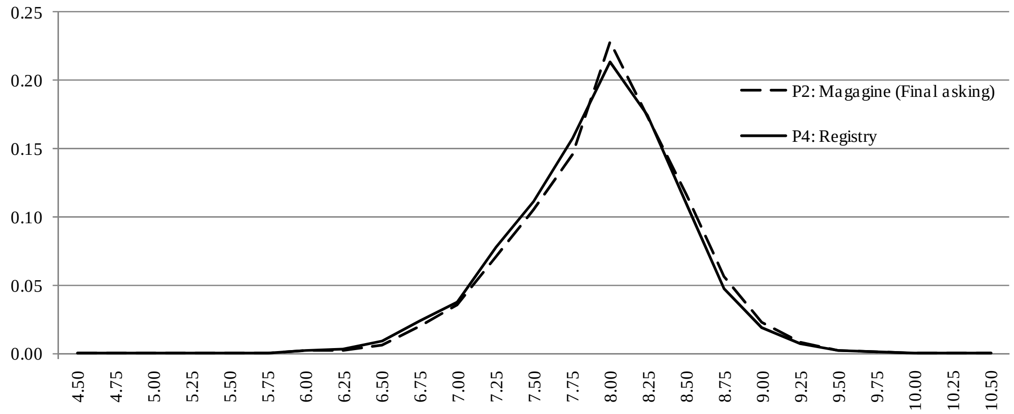


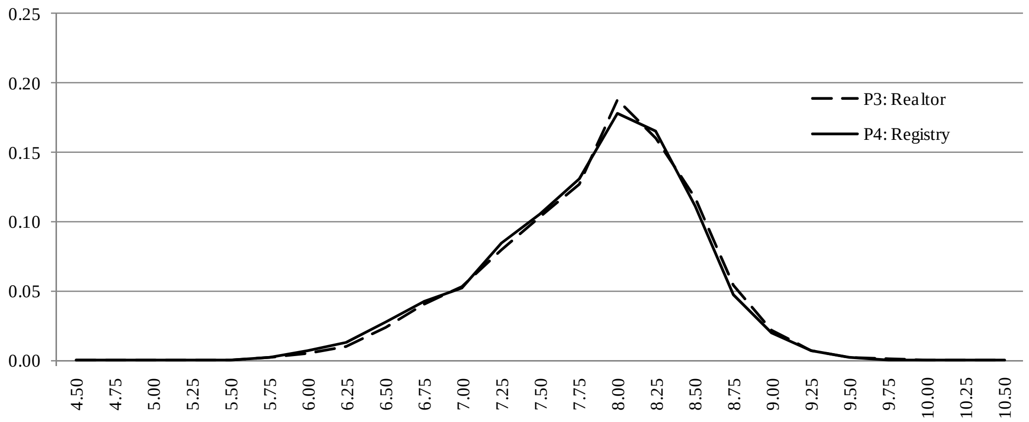

**Figure 5: Densities for relative prices**




Densities for $P_1$ and $P_4$

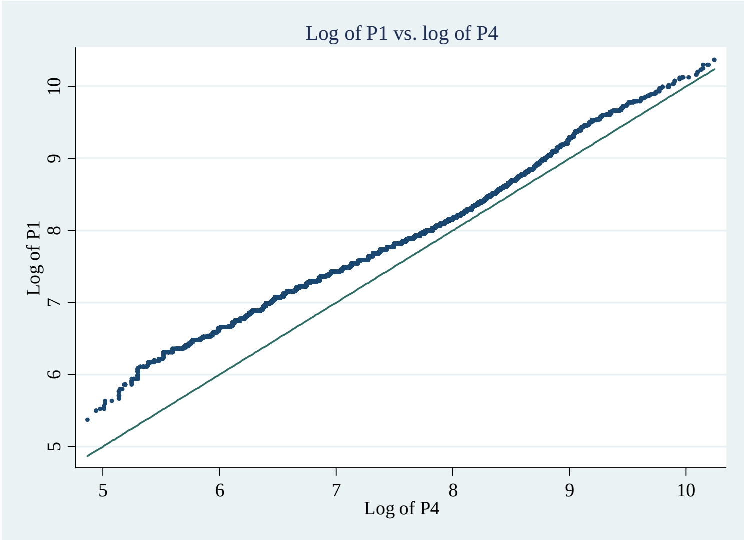


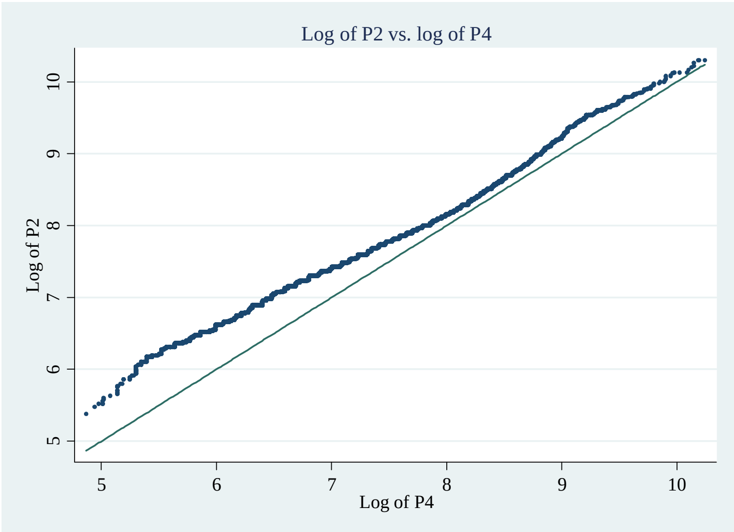

Densities for $P_2$ and $P_4$




Densities for $P_3$ and $P_4$

**Figure 6: Price densities for housing units observed in two datasets**



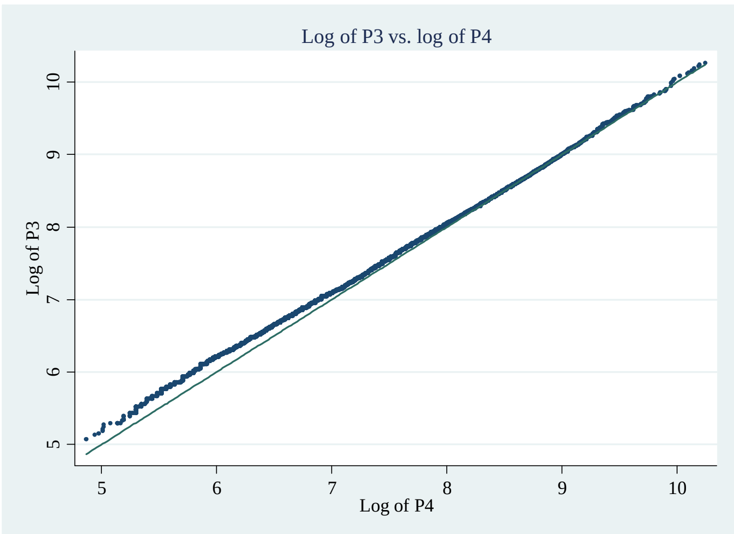
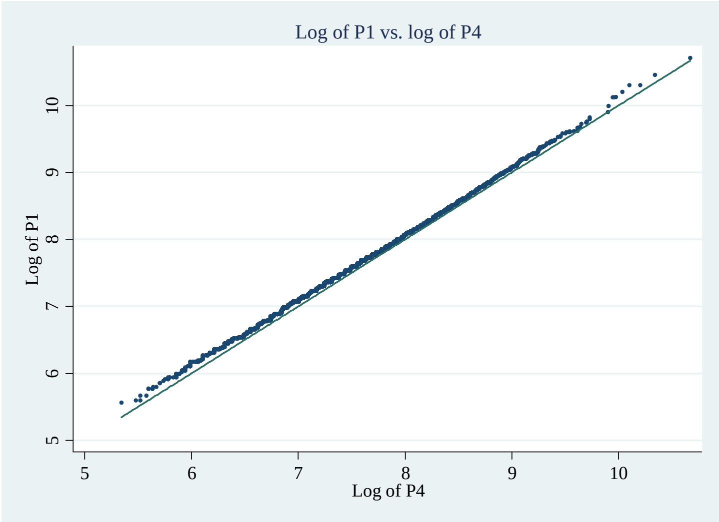
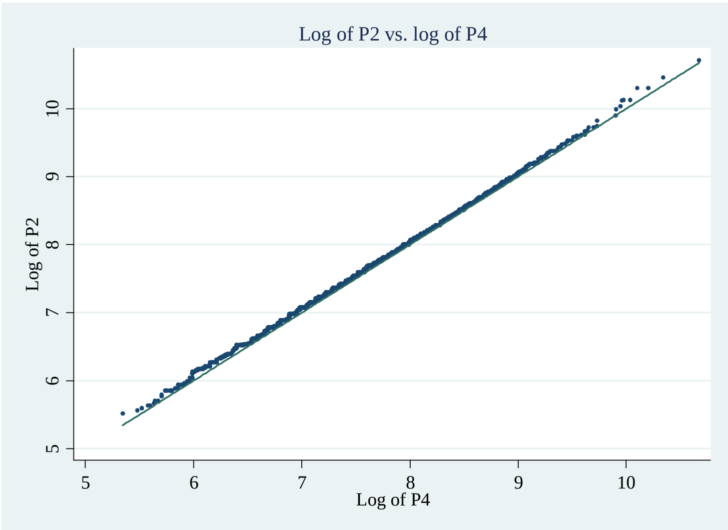




**Figure 7(a): Quantile-quantile plots for raw data**



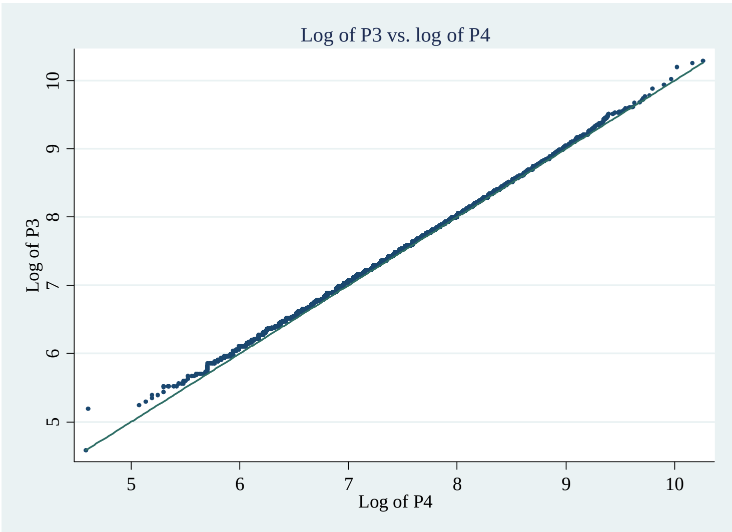
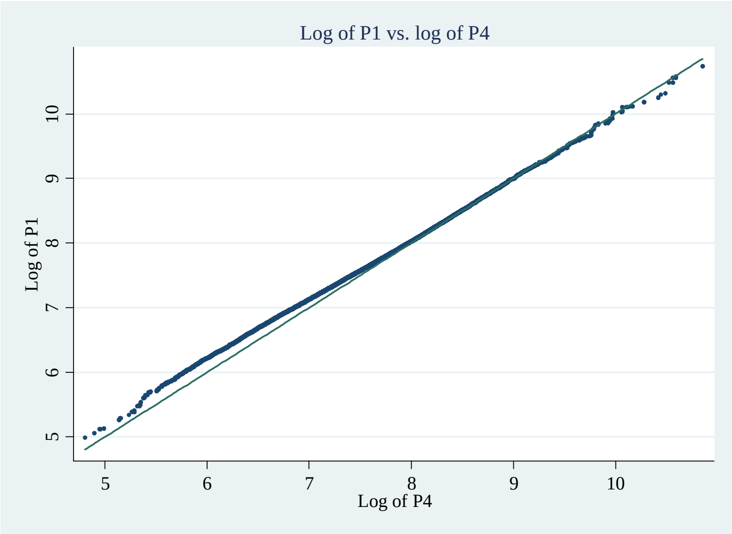
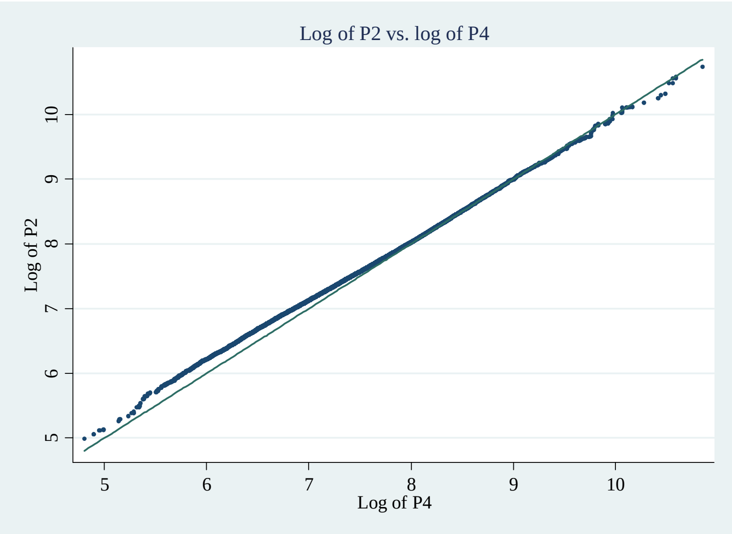




**Figure 7(b): Quantile-quantile plots for quality adjusted prices by intersection approach**



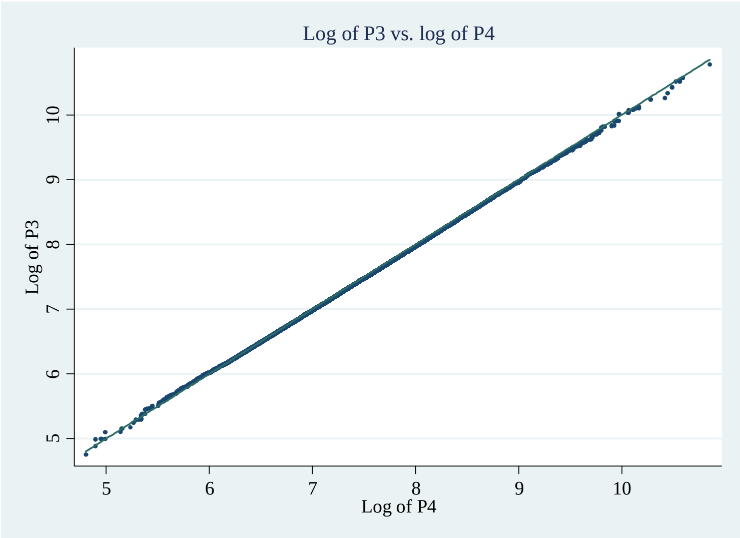




**Figure 7(c): Quantile-quantile plots for quality adjusted prices by quantile hedonic approach**

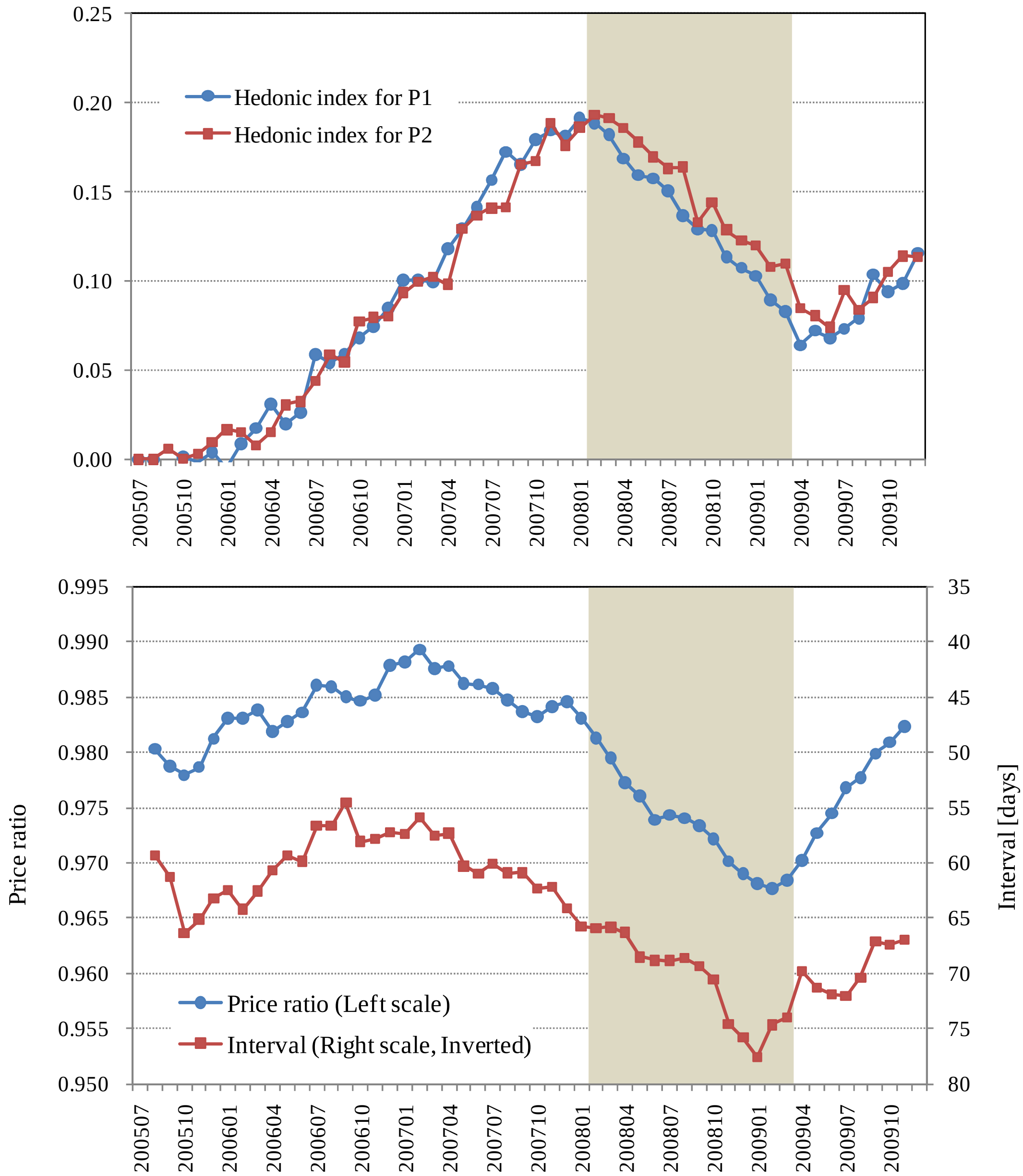
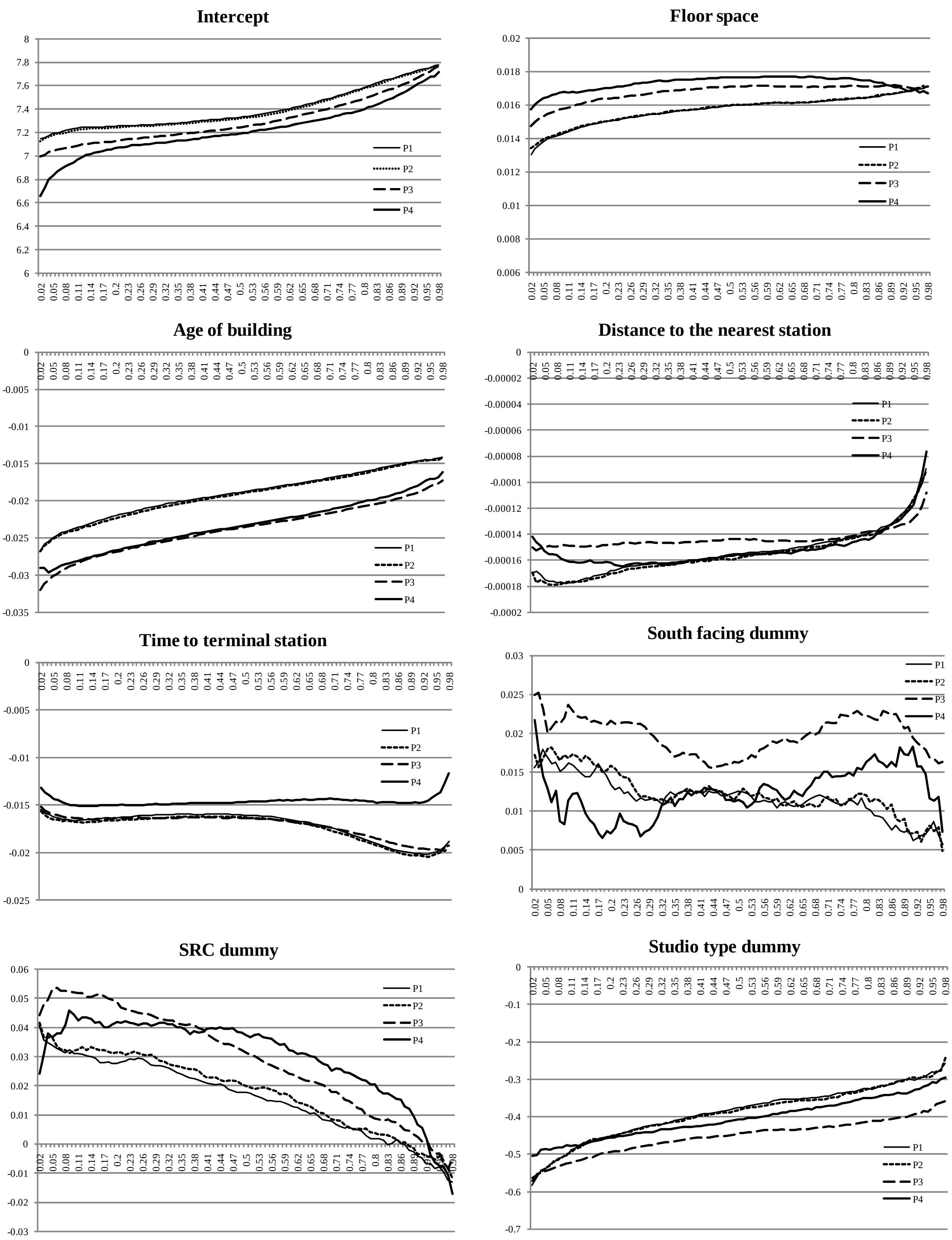
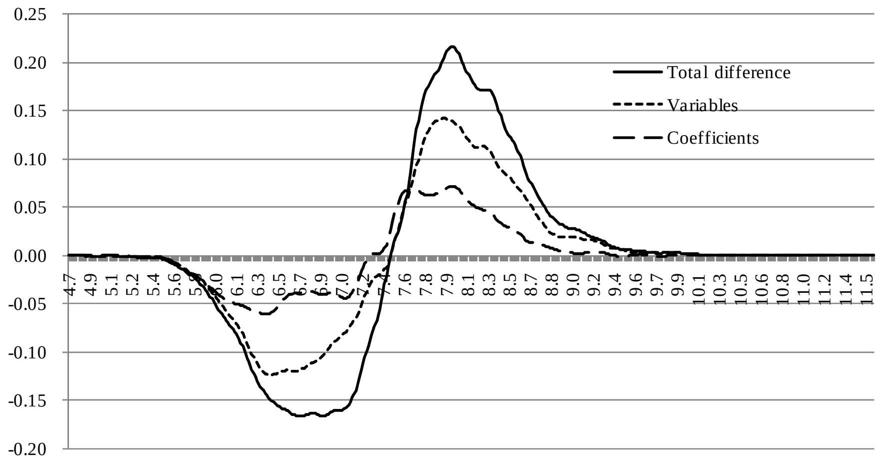


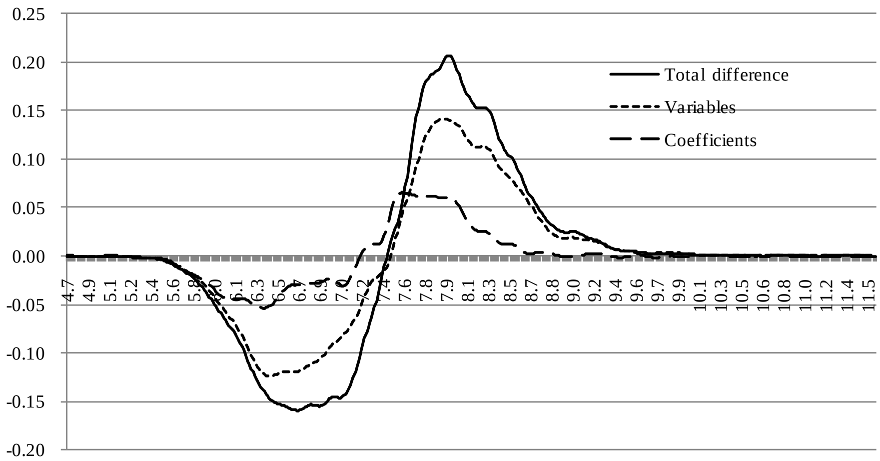

**Figure 8: Fluctuations in the price ratio and the interval for $P_1$ and $P_2$**



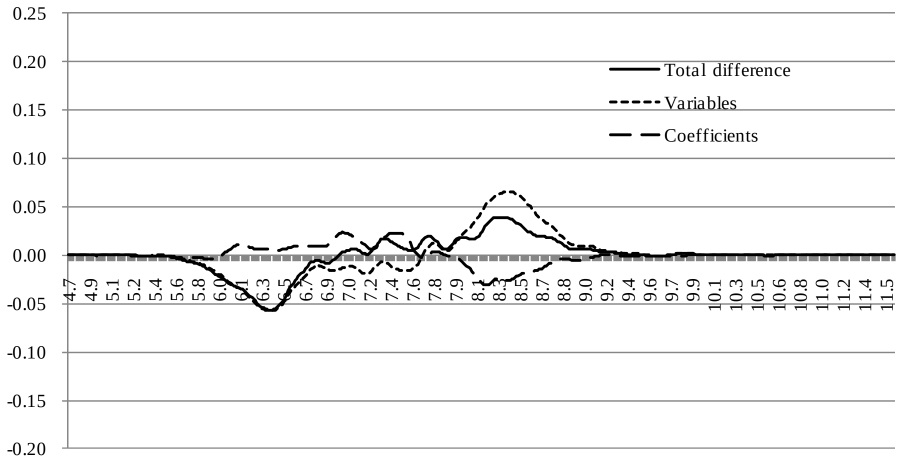

# Figure 9: Quantile regression coefficients




Difference between $P_1$ and $P_4$




Difference between $P_2$ and $P_4$




Difference between $P_3$ and $P_4$

**Figure 10: Decomposition of density differences**