Replicating Japan's CPI Using Scanner Data

Satoshi Imai* Tsutomu Watanabe†

June 26, 2015

Abstract

We examine how precisely one can reproduce the CPI constructed based on price surveys using scanner data. Specifically, we closely follow the procedure adopted by the Statistics Bureau of Japan when we sample outlets, products, and prices from our scanner data and aggregate them to construct a scanner data-based price index. We show that the following holds the key to precise replication of the CPI. First, the scanner data-based index crucially depends on how often one replaces the products sampled. The scanner data index shows a substantial deviation from the actual CPI when one chooses a value for the parameter associated with product replacement such that replacement occurs frequently, but the deviation becomes much smaller if one picks a parameter value such that product replacement occurs only infrequently. Second, even when products are replaced only infrequently, the scanner data index differs significantly from the actual CPI in terms of volatility. The standard deviation of the scanner data-based monthly inflation rate is 1.54 percent, which is more than three times as large as that for actual CPI inflation. We decompose the difference in volatility between the two indexes into various factors, showing that it mainly stems from the difference in price rigidity for individual products. We propose a filtering technique to make individual prices in the scanner data stickier, thereby making scanner data-based inflation less volatile.

Keywords: consumer price index; scanner data; sampling; price rigidity; menu costs; sale and regular prices

*Corresponding author: Satoshi Imai. Statistics Bureau of Japan. E-mail: s2.imai@soumu.go.jp. This paper was prepared for the Ottawa Group meeting held on May 20-22, 2015, Tokyo. This is a preliminary version, so please do not quote without permission from the authors. We would like to thank Erwin Diewert, David Fenwick, Jan de Haan, Shun-ichi Shimakita, Sei Ueda, Kota Watanabe, and participants at the Ottawa Group 2015 for useful comments and discussions. This research forms part of the project on "Understanding Persistent Deflation in Japan" funded by a JSPS Grant-in-Aid for Scientific Research (No. 24223003).

†Graduate School of Economics, University of Tokyo. E-mail: watanabe@e.u-tokyo.ac.jp. Website: https://sites.google.com/site/twatanabelab/

1 Introduction

Scanner data has started to be used by national statistical offices in a number of countries, including Australia, the Netherlands, Norway, Sweden, and Switzerland, for at least part of the production of their consumer price indexes (CPIs). Many other national statistical offices have also already started preparing for the use of scanner data in constructing their CPIs. The purpose of this paper is to empirically examine whether price indexes based on scanner data is consistent with price indexes constructed using the traditional survey based method.

The consistency between the old and new indexes is an important issue for national statistical offices, which are responsible for making sure that statistical properties of the CPI remain unchanged even when switching to scanner data.[1] If the new and old indexes are not consistent with each other, the central bank, for example, will not be able to conduct monetary policies, such as inflation targeting, based on the new index. However, there are several reasons to believe that the two indexes may have different time series properties.[2]

The first potential source of discrepancy is related to differences between offer prices and transaction prices. In the traditional survey method adopted by almost all countries including Japan, price collectors collect price tag information. Such information may not be the same as the information contained in scanner data, which is based on actual transaction prices. As shown by the literature on asset pricing in financial markets, quoted prices, such as prices quoted by stock market dealers, have different time series properties than transaction prices in terms of volatility, serial correlation, cross sectional correlation, and so on.[3] Similar differences may exist for goods and service price series. Consider, for example, a situation in which there is no transaction record for a particular product at a particular outlet. In this

[1] Fenwick (2014) compares the advantages and disadvantages of CPI construction based on scanner data relative to the current method based on price surveys.

[2] Note that the consistency between scanner data-based price indexes and CPIs is an important issue not only for national statistical offices but also for the private sector, especially market participants in stock, bond, and foreign exchange markets. Market participants want to forecast future policy, such as the policy rate set by the central bank, several months ahead and therefore closely monitor the official CPI, which is regarded as an important determinant of future policy. However, the official CPI is released only once a month, with more than a one-month lag. Given this limitation, market participants have recently started to use information on prices from alternative sources such as scanner data and online price data. For example, the Billion Prices Project at the Massachusetts Institute of Technology has been releasing daily inflation data since 2011, which are constructed based on online price data collected through web scraping. Similarly, the UTokyo Daily Price Project at the University of Tokyo has been releasing daily inflation data since May 2013, which is based on scanner data. Daily inflation data released by the two price projects is widely used by market participants who want to predict the CPI. Given this, it is widely regarded as important by market participants to empirically assess whether and under what circumstance the daily inflation figure is a good leading indicator of the official CPI.

[3] See, for example, Hautsch (2004).

case, transaction price information is missing but price tag information for that product is still available, which may produce non-negligible differences between the two price indexes in terms of their time series properties. Watanabe and Watanabe (2014), for instance, show that missing observations in the scanner data tend to occur on the day immediately after a temporary price reduction ends. They argue that such missing observations likely create chain drift in scanner data-based price indexes.

Second, discrepancies between the two indexes may arise due to differences in sampling procedures. In the current method based on price surveys, the replacement of outlets and products to be sampled is decided by national statistical offices, although detailed information on the criteria actually adopted for replacement may not necessarily be disclosed. On the other hand, scanner data provides information on the number of customer visits for outlets and on the quantities sold for products, so that it is possible to use this information as a criterion when deciding which outlets and products to sample. For example, Imai et al. (2013) choose the set of outlets based on the number of customer visits to the outlets and the set of products based on the quantity sold when they construct a scanner data-based price index for Japan. However, there is no guarantee that the set of outlets and the set of products chosen this way coincide with those chosen by the national statistical offices.

Third, discrepancies between the two indexes may arise due to differences in the way temporary price reductions are dealt with. In most countries including Japan, price collectors are instructed to collect regular prices. In the case of Japan, price collectors are instructed to exclude "extra-low prices due to bargain, clearance, or discount sales, and quoted for less than eight days" (Statistics Bureau of Japan (2013)). Therefore, when a particular product is on sale at a particular outlet, price collectors are instructed to get information on its regular price from the outlet manager. To replicate the CPI methodology, it is necessary to replace sale prices with regular prices. Specifically, one can apply various filtering techniques, such as those empirically examined by Chahrour (2011), to scanner data to estimate regular prices. However, it is likely that regular prices estimated in this way may differ from regular prices obtained by price collectors.

The aim of this paper is to empirically examine the consistency between scanner data- and traditional survey-based price indexes and, if the former are found to be inconsistent with the latter, to investigate why this is so and propose remedies. Specifically, we construct a scanner data-based price index for Tokyo closely following the procedure currently adopted by the Statistic Bureau of Japan (JSB). The current JSB procedure for product sampling is based

on purposive sampling in which product type specifications for each of the item categories are defined in advance and products are sampled only from a set of candidate products with these specifications. As for the aggregation procedure, the JSB takes the unweighted arithmetic mean of prices of products belonging to an item category (i.e., the JSB constructs a Dutot index) for lower level aggregation, and applies fixed weight Laspeyres weighting for upper level aggregation. The sampling and aggregation procedure adopted by the JSB is not very different from the procedures adopted by statistical offices in most other industrial countries, although it does differ substantially from procedure employed in the United States, where random sampling rather than purposive sampling is used for product sampling.

The main findings of the paper are as follows. First, we show that the estimated scanner data-based price index crucially depends on how frequently products to be sampled are replaced. Specifically, the root mean squared error between the scanner data-based price index and the actual CPI is not negligible when we choose the value of the product replacement parameter such that replacement occurs very frequently. However, it becomes smaller when we pick a value for this parameter so that replacement occurs only infrequently. The result means that replacing products frequently in the scanner data-based index does not yield a good approximation to the CPI based on current JSB practices. Instead, the JSB practices are characterized by a degree of inertia in the sense that products belonging to the current sample are not replaced quickly even if their quantity sold declines.

Second, we show that the estimated scanner-based index has much higher volatility than the actual CPI, even if inertia in product sampling is incorporated. Specifically, the standard deviation for the month-on-month inflation rate is 0.57 percent for the scanner data-based index, while it is 0.41 percent for the actual CPI. We decompose the difference in volatility between the two indexes into various factors, which shows that it mainly stems from differences in the frequency of price adjustments for individual products. That is, actual CPI inflation is less volatile since individual prices in the CPI data are stickier.

Third, we show that small price changes, such as price changes of less than ± 4 percent, are less likely to occur in the CPI data. Together with the second finding, this suggests that menu costs (i.e., transaction costs associated with price changes) play a more important role in the CPI data. We propose and implement a filtering technique to make scanner data-based inflation less volatile.

The rest of the paper is organized as follows. Section 2 explains the methodology and the datasets we use in the paper. Section 3 presents our main empirical results. Section 4

concludes the paper.

2 Methodology and Data

In this section, we describe the procedures used by the JSB to construct the CPI and explain how we replicate their procedures. Specifically, we outline the JSB's outlet sampling, product sampling, and price sampling procedures (Section 2.1) and upper and lower-level aggregation procedures (Section 2.2). In addition, we provide a description of the data used in the paper (Section 2.3).

2.1 The JSB's sampling methodology

CPIs in different countries are constructed following a set of common rules, which are described in various documents such as ILO (2004). Nevertheless, there still remain several important methodological differences, one of which is differences in product sampling, with some countries employing purposive and others random sampling. In purposive sampling, the statistical office of a country defines product type specifications for each of the item categories. Products are sampled only from a set of candidate products with these specifications. On the other hand, in random sampling, products are randomly chosen among all products belonging to an item category (i.e., without specifying a set of candidate products).

From a statistical perspective, purposive sampling has some undesirable characteristics, including sampling bias (i.e., the prices of sampled products may not come from the true price distribution) and lower sampling efficiency (i.e., the variance of prices of sampled products may be larger than the corresponding variance in the case of random sampling). However, purposive sampling has advantages from a practical perspective in that the process of narrowing the range of candidate products makes sampled products more homogeneous, thereby making estimated price indexes less volatile even in the case of high product substitution. As a result, many countries, including Japan, have adopted purposive sampling, while only a limited number of countries, including the United States, have adopted random sampling.

In this paper, we employ purposive sampling to collect prices from the universe of scanner data. The purposive sampling conducted in this paper is based on the list of product types, with product type specifications, used by the JSB, which we refer to as the JSB product type specifications.[4] In the rest of this subsection, we explain how we mimic the JSB methodology in terms of outlet sampling, product sampling, and price sampling.

[4] A complete list of product type specifications is available in Statistics Bureau of Japan (2012).

Outlet sampling The number of outlets to be sampled differs from item to item and changes over time. For example, since January 2010, the number of outlets for "wheat flour" has been 42 and that for "butter" 12. This means that, for wheat flour, price collectors collect 42 prices from 42 outlets located in Tokyo. The number of outlets for each item is determined based on various factors, including the price dispersion across outlets, and released by the JSB. The number of outlets to which we send our virtual price collectors is the same as the number of outlets to which the JSB actually sends price collectors.

According to the JSB, outlets to be visited by price collectors are chosen based on how representative that outlet is in that area. Following this, we choose outlets based on the number of customer visits to an outlet. Specifically, at the start of our sample period (January 2000), we choose outlets based on the number of customer visits to an outlet over the last one month. As for the subsequent months, we randomly choose outlets based on the following procedure. We take outlet i, which is already included in the sample, and outlet j, which is not included in the sample but a candidate to be included, and calculate the following replacement probability:

$$\Pr\left(\text{Outlet replacement}\right) = \left[1 + \left(k_S^{-1}\frac{n_j^S}{n_i^S}\right)^{-1}\right]^{-1} \tag{1}$$

where n_i^S and n_j^S represent the number of customer visits to outlets i and j over the last one month, and k_S is a parameter that takes a value equal to or greater than unity. This procedure is repeated for all pairs of outlets until we finally end up with the list of outlets to be sampled.

Eq (1) states that the probability of outlet replacement is higher the larger is n_j^S relative to n_i^S. However, the parameter k_S also affects the probability of outlet replacement; that is, if k_S is very large, outlet replacement is less likely to occur even if n_j^S is large relative to n_i^S. In this sense, k_S is a parameter governing the frequency/infrequency of outlet replacements. In our replication exercise, we search the optimal value of k_S such that the discrepancy between the scanner data-based price index and the actual CPI is minimized. Note that the maximum number of outlets in the scanner data is around 30, while the JSB for some items samples up to 42 outlets, so that for certain items the scanner data does not have the required number of outlets. In this case, we choose all outlets for these items.

Product sampling Once an outlet is picked, we then choose a product out of the set of products that meet the JSB specifications, based on the quantity sold at that outlet over the

last one month. Let us explain how we specify the set of candidate products, taking butter as an example. According to the JSB list of product types, the product type specifications for butter are as follows:

JSB Product Type Specifications for Butter

Jul. 1996 - Jan. 2001	"Snow Brand Hokkaido Butter"
Jan. 2001 - present	200g. Packed in a paper container. Excluding unsalted butter.

Note that only a single product, "Snow Brand Hokkaido Butter," was on the list from July 1996 to January 2001, while multiple products were allowed in the more recent period. Based on this information, we produce a list of product barcodes - called JAN (Japan Article Number) codes - that meet the JSB product type specifications. Our task is very simple for the period from January 2000 to January 2001: we just look for the unique JAN code corresponding to "Snow Brand Hokkaido Butter." On the other hand, for the period from February 2001 to the end of our sample period, we look for the JAN codes of products that meet the specifications described above. Specifically, we do so using supplementary information on each JAN code, including the name of a product, brand, model number, net quantity, and ingredients. This process is done by using a text matching technique ("regular expression"). We find that the number of products (i.e., the number of JAN codes) that meet the above specifications is 31. Among these 31 products, we choose a single product based on the quantities sold over the last one month at a particular outlet chosen through the outlet sampling procedure described above.[5]

The procedure to choose a product from among many candidates is similar to the one we adopt for outlet sampling. We take product i, which is already included in the sample, and product j, which is not included in the sample but a candidate to be included, and calculate the replacement probability as follows:

$$\Pr\left(\text{Product replacement}\right) = \left[1 + \left(k_P^{-1}\frac{n_j^P}{n_i^P}\right)^{-1}\right]^{-1} \tag{2}$$

where n_i^P and n_j^Prepresent the number of quantities sold for product i and j over the last one month at a particular outlet chosen through the outlet sampling procedure, and k_P is

[5]Note that in the example given here, "unsalted butter" can be regarded as a negative characteristic in the sense that products with that characteristic ("unsalted butter") are *excluded* from the product specification, while "200g" and "packed in a paper container" can be seen as positive characteristics. For butter, the number of products based on the full range of characteristics (including the negative characteristic "excluding unsalted butter") is 31, as mentioned above, while the number of products based on positive characteristics only (i.e., "unsalted butter" is not excluded) is 123.

the parameter governing the frequency/infrequency of product replacements, which takes a value equal to or greater than unity. Note that a larger k_P means that product replacement is less likely to occur even if n_j^P is large relative to n_i^P.

Table 1, which is taken from Imai et al. (2013), presents the number of products that meet the JSB product type specifications.[6] For example, the total number of products (i.e., the number of JAN codes) for item code 1321 ("Butter") is 369, and the number of products that meet the JSB product type specifications is 31. The share of products that meet the JSB specifications is very small (8 percent), although the sales share of those products is relatively large (45 percent). The number of products belonging to all of the item categories covered by our scanner data is 462,906, among which 70,966 products meet the JSB specifications (15 percent).

Quality adjustment When product replacement takes place, the new and old products may differ in quality, so that they are not directly comparable. To make them comparable, the JSB applies a quality adjustment procedure to the new and old products. Specifically, the JSB employs three different methods for quality adjustment; (1) direct comparison, (2) direct quality adjustment, and (3) imputation.[7]

Direct comparison is employed when the new and old products are essentially the same. In this case, the price of the new product and the price of the old product are treated as if no product replacement occurred. On the other hand, direct quality adjustment is employed when information about the change in quality between the old and new products is available. For example, if the old and new products differ only in terms of their quantity, and prices can be regarded to depend linearly on product quantity, the price of the new product is adjusted using the quantity ratio between the old and new products (this is referred to as the "quantity-ratio method" by the JSB). More generally, if information on product characteristics is available for the old and new products, a hedonic regression is applied to estimate quality adjusted prices. Another way to conduct direct quality adjustment is to use information on the observed price difference between the old and new products at a particular point in time. Specifically, if prices of the new and old products are available in months t and $t-1$ and it is safe to assume that the price difference between them reflects the quality

[6] Note that the JSB list of product type specifications is updated every five years, although for some items minor modifications are made more frequently.

[7] For more on the JSB's quality adjustment procedure, see `http://www.stat.go.jp/english/data/cpi/1586.htm`.

difference between the old and new products, the price difference between the old and new products in $t-1$ is regarded as a measure of the quality difference and used to estimate the quality adjusted price of the new product in t (this is referred to as the "sample overlap method" by the JSB).

Finally, imputation is employed when neither information on product characteristics nor information on prices in $t-1$ and t is available. In this case, an estimate of constant-quality price change is made by imputation. Specifically, based on the assumption that the price change for the new product from $t-1$ to t is the same as price changes for the other products in the same item category, an estimate of the price of the old product in t is computed by multiplying the price of the old product in $t-1$ by the rate of inflation between $t-1$ and t for the other products belonging to the same item category.

To replicate the JSB quality adjustment procedure, we need detailed information on the JSB procedure, including under what circumstances which quality adjustment method should be adopted. However, detailed information on the JSB 's quality adjustment procedure, including which method was adopted for each of the product replacements that actually occurred, is disclosed to the public only once a year with a substantial time lag. Given this limited information, we take the following approach. We put first priority to the quantity-ratio method. We employ this method whenever information on the weight or size of the new and old products is available. Otherwise, however, we employ the imputation method. We do not use the other quality adjustment methods employed by the JSB. Needless to say, this may not be a good approximation to the JSB's practice. To see whether this is a good approximation or not, we will conduct an experiment in the next section by applying our quality adjustment procedure to the CPI source data (i.e., the individual price data actually collected by JSB price collectors) rather than the scanner data and construct a price index. This experiment uses the same price data actually used in the CPI, but it differs from the actual JSB procedure in terms of quality adjustment and lower and upper aggregation. Through this exercise, we can learn whether our quality adjustment and aggregation procedures are a good approximation to the JSB's practices.[8]

[8]Note that when outlet replacement occurs, we need to conduct quality adjustment to products sold at the new and old outlets. Again, however, we do not have much information on how this is handled by the JSB. In the empirical exercise we conduct in the next section, we will extend the imputation method so that it can be applied to outlet replacements.

Price sampling Price collectors are instructed by the JSB not to collect sale prices. Specifically, price collectors are instructed to exclude "extra-low prices due to bargain, clearance, or discount sales, and quoted for less than eight days" (Statistics Bureau of Japan (2013)). To mimic this practice, we treat temporary price reductions as follows. First, we define a temporary price reduction as a price reduction where the price goes back to its original level. Next, we then identify such temporary price reductions for each product at each outlet. If the duration of a temporary price reduction is equal to or more than eight days, we do not apply any special treatment; however, if it is less than eight days, we do not use that price and instead look for the "regular" price. Specifically, we assume that the regular price is equal to the price level just before the temporary price reduction.

As for the timing of price collection, we follow the current practice adopted by the JSB. That is, price collectors are instructed to collect prices on either Wednesday, Thursday, or Friday of the week which includes the 12th of the month. The order of priority regarding the three days is Thursday, Wednesday, and Friday. If no transaction is recorded during these three days for a particular product in a particular month, we search for a record of transactions retroactively from that date to the 1st of that month.

Note that price collectors may fail to collect the price of a product at an outlet if the product was temporarily unavailable at the outlet. This is referred to as "missing prices" by the JSB. Again we do not have much information as to how this is handled by the JSB, but in the United States, according to Bureau of Labor Statistics (2007), missing prices are estimated using "cell-relative imputation." We will follow this procedure here. Specifically, based on the assumption that the price change from $t-1$ to t for the product whose price is missing is the same as price changes for the other products in the same item category, an estimate of the price of the product in t is computed by multiplying the price of the product in $t-1$ by the rate of inflation between $t-1$ and t for the other products belonging to the same item category.

2.2 Aggregation at lower and upper levels

Aggregation at the lower level For aggregation at the lower level, we again follow the JSB procedure by employing the unweighted arithmetic mean of price levels across product-outlet combinations (i.e., the Dutot index). That is, the price index for item i in region r in

month t, $P_{r,i}(t)$, is defined as

$$P_{r,i}(t) \equiv \left(\frac{1}{n}\right) \sum_{(o,j)\in A_{r,i}} P_{r,i,o,j}(t) \tag{3}$$

where $P_{r,i,o,j}(t)$ represents the price in month t of product j, which belongs to item i, quoted at outlet o located in region r, n is the number of products collected for an item in a region, and $A_{r,i}$ is the set of product-outlet combinations obtained through the process of outlet and product sampling explained earlier.

Aggregation at the upper level Next, we construct a fixed-base Laspeyres index by aggregating the lower level indexes. The price index in region r is defined as

$$I_r(t) \equiv \sum_i \omega_{r,i} \frac{P_{r,i}(t)}{P_{r,i}(t_0)} \tag{4}$$

where $\omega_{r,i}$ is the consumption weight for item i in region r in the base year ($t = t_0$), satisfying $\sum_i \omega_{r,i} = 1$. The weight $\omega_{r,i}$ is taken from the *Family Income and Expenditure Survey* conducted by the Japanese government. [Finally, we construct the price index for Tokyo as a whole by aggregating the regional indexes:

$$I(t) \equiv \sum_r \omega_r I_r(t) \tag{5}$$

where ω_r represents the consumption weight for region r with $\sum_r \omega_r = 1$.

2.3 Data

We use two datasets: one is the CPI source data compiled by the JSB, while the other is the scanner data compiled jointly by the UTokyo Price Project and Nikkei Digital Media Inc.[9] The CPI source data is available from January 2010 to July 2014, while the scanner data is available from January 2000 to July 2014.

Our scanner data contains price and quantity information at a daily frequency for more than 300,000 products sold at about 300 supermarkets in Japan. The products consist mainly of food, beverages, and other domestic nondurables (such as detergent, facial tissue, shampoo, soap, and toothbrushes). Outlet coverage is relatively high for large cities such as Tokyo but not for other areas. In this paper, we restrict our attention to Tokyo. Table 2 shows the

[9]The CPI source data is not made available to the public. One of the authors of this paper, Satoshi Imai, is a staff member of the JSB and conducted the analyses using this data in this capacity.

number of outlets, products, and observations in Tokyo. The number of outlets in Tokyo changes over time but is around 30. For the year 2013, the number of outlets is 32, the number of products is 151,000, and the total number of observations is 50 million. Compared to the entire dataset, the data for Tokyo accounts for about 40 percent in terms of the number of products and about 11 percent in terms of the number of observations.

Table 3 shows the turnover of products sold at the 14 outlets in Tokyo that existed throughout the entire sample period. More than 30,000 new products were introduced each year and about the same number of products exited from the market. Product turnover hit a peak in 2008, when product replacements tended to be accompanied by product downsizing allowing firms to raise effective prices without changing nominal prices (see Imai and Watanabe (2013) for more on this). The ratio of new products relative to existing products was about 35 percent, while the corresponding exit rate was about 34 percent, both of which are higher than for the entire dataset covering Japan.

The CPI dataset we use in this paper covers the Tokyo area and contains 94 items out of the 588 items, accounting for 8.1 percent in terms of CPI weights. The items included in our dataset are non-fresh foods, processed food, and daily necessities. These items are chosen so that they are also included in the scanner data. Table 4 presents a list of the 94 items and their CPI weight. The sample period is January 2010 to July 2014. The number of individual price observations collected for the 94 items is about 2,800 each month, and the number of price observations over the entire sample period is 151,525.

3 Empirical Results

3.1 Replicating the CPI using the CPI source data

The first exercise we conduct is to apply our methodology to the CPI source data (i.e., the set of prices actually collected by JSB price collectors). In this exercise, we do not use our outlet, product, and price sampling procedure and instead use the prices already collected by JSB collectors, so that any differences between our price index and the actual CPI do not stem from the sampling procedure but from the procedure for quality adjustment and/or lower and upper level aggregation. In this way, we can check whether our quality adjustment and aggregation procedures are a good approximation to the procedure actually adopted by the JSB.

The result is presented in Figure 1, with the price level shown in the upper panel and

the year-on-year inflation rate in the lower panel. The figure shows that we almost perfectly reproduce the actual CPI in terms of both the price level and price changes, except in mid-2013, when our estimate for the price level is slightly lower than the actual CPI. Note that we use only two types of quality adjustment procedure (i.e., the imputation method and the quantity-ratio method) when product replacement occurs, while the JSB may have employed other procedures as well, as explained in the previous section. The difference in mid-2013 therefore may be due to the JSB employing these other quality adjustment procedures for product replacements that occurred during that period. However, the difference in terms of the price level in mid-2013 is 0.3 percent at most, so it is negligible. We can conclude from this exercise that our quality adjustment and aggregation procedures are a very good approximation to the JSB procedures.

3.2 Replicating the CPI using the scanner data

Next, we repeat the same exercise but now use our outlet, product, and price sampling procedures. The result is shown in Figure 2. We use the probabilities given by eqs (1) and (2) when conducting outlet and product sampling. The values for k_S and k_P in (1) and (2) are set to $k_S = 1$ and $k_P = 1$. This means that we choose parameter values such that outlet and product replacement occur frequently.

We repeat probabilistic sampling based on eqs (1) and (2) 200 times, and the mean of the 200 outcomes is represented by the red line in Figure 2. The standard deviation of the 200 outcomes is also calculated to estimate the confidence interval (i.e., the mean $\pm$ one standard deviation), which is represented by the shaded area surrounding the red line. Figure 2 shows the following. First, as seen in the lower panel, the scanner data-based inflation tends to be below actual CPI inflation during the first half of the sample period (i.e., 2001-2007). The difference is not negligible and at times reaches 2 percentage points. Second, scanner based inflation is much more volatile than actual CPI inflation. Scanner based inflation exhibits particularly large fluctuations in 2004-2005 and 2011-2012, while actual CPI inflation is not that volatile, even during these periods. Third, the difference between the two indexes has increased since April 2014 when the consumption tax rate was raised from 5 percent to 8 percent. Specifically, actual CPI inflation has started to decline since April 2014, but scanner data based inflation continues to rise, reaching 4 percent in July 2014, the last month in our dataset.

Where do these differences come from? There are several possibilities but one potential

source is the difference in sampling procedures. The number of outlet replacements over the entire sample period is 3,000 for the scanner data-based index, while it is 1,340 for the actual CPI, indicating that outlet replacements are more likely to occur in the scanner data-based index. Similarly, the number of product replacements is 4,730 for the scanner data-based index and 2,793 for the actual CPI, again indicating that product replacements are much more likely to occur in the scanner data-based index.

To investigate this further, we repeat the exercise using different values for k_S and k_P. We then calculate the root mean squared error (RMSE) for the discrepancy between scanner data-based inflation and actual CPI inflation for different pairs of k_S and k_P. The result is presented in Table 5 and Figure 3. Note that, for a particular pair of k_S and k_P, we repeat probabilistic sampling based on eqns (1) and (2) 200 times and calculate the RMSE 200 times. The number shown in Table 5 is the average of the 200 RMSEs obtained this way.

As can be seen in the table, the RMSE does not change much for different values of k_S, suggesting that the outcome does not depend much on the frequency of outlet sampling. However, the RMSE becomes smaller as k_P increases, indicating that a lower frequency of product replacements makes the outcome more similar to the actual CPI. In particular, as shown in Figure 3, the decline in the RMSE is substantial when k_P increases from 1 to 5, although it is much less pronounced when k_P increases beyond that. This result indicates that assuming infrequent product replacement, with a k_P equal to 5 or slightly higher, provides a good approximation to the JSB procedure.

Figure 4 shows the outcome obtained in the case of $k_S = 5$ and $k_P = 10$. Compared to the outcome with $k_S = 1$ and $k_P = 1$, which is presented in Figure 2, we see considerable improvement. In particular, scanner data-based inflation in 2001-2007 is now much closer to actual CPI inflation. However, we do not see a great reduction in volatility differences: scanner data-based inflation still exhibits much higher volatility than actual CPI inflation. Also, we still have a non-negligible deviation of scanner data-based inflation from actual CPI inflation since the consumption tax increase in April 2014.

Finally, Figure 5 presents an item-by-item comparison of the two indexes, with the horizontal axis representing the item codes (see Table 3 for item names) and the vertical axis showing the mean and standard deviation of month-on-month inflation. The sample average of month-on-month price changes for actual CPI inflation is represented by the dotted blue line and that for scanner data-based inflation with $k_S = 5$ and $k_P = 10$ by the dotted red line. We see no substantial differences between the two indexes in terms of the average rate

of inflation, indicating again that infrequent replacement with $k_S = 5$ and $k_P = 10$ is a good approximation to the JSB procedure. However, turning to the standard deviation of month-on-month price changes, which is shown by the blue solid line for actual CPI inflation and by the red solid line for scanner data-based inflation, we see that the standard deviation for scanner data-based inflation is substantially higher than that for actual CPI inflation for almost all items and in fact more than twice as high for some items.

3.3 Why is scanner data-based inflation more volatile?

Why is scanner data-based inflation more volatile than actual CPI inflation? In this subsection, we search for the causes and then propose a method to reduce the volatility of scanner data-based inflation to a level similar to that of CPI inflation.

3.3.1 Decomposition of inflation volatility into intensive and extensive margins

As a first step to investigate the causes of the high volatility in scanner data-based inflation, we decompose inflation volatility into several components. Let us denote the month-on-month inflation for item c in month t by π_{ct}, the fraction of products in item c that experience price changes in month t by Fr_{ct}, and the average size of price changes for those products in item c that experience price changes in month t by dP_{ct}. The monthly inflation rate in t can be decomposed into the fraction of products that experience price changes in t and the average size of price changes for these products. That is,

$$\pi_{ct} \equiv \mathrm{Fr}_{ct} \times dP_{ct} \tag{6}$$

We then take the variance of the first-order Taylor series expansion of eq (6) around the means of Fr_{ct} and dP_{ct} (i.e., $E(\mathrm{Fr}_{ct})$ and $E(dP_{ct})$) to arrive at the following equation:

$$Var(\pi_{ct}) = \underbrace{Var(dP_{ct})\,[E(\mathrm{Fr}_{ct})]^2}_{\text{Intensive margin term}} + \underbrace{Var(\mathrm{Fr}_{ct})\,[E(dP_{ct})]^2}_{\text{Extensive margin term}} + \text{Other terms} \tag{7}$$

where $Var(\pi_{ct})$, $Var(dP_{ct})$, and $Var(\mathrm{Fr}_{ct})$ are the time series variances of π_{ct}, dP_{ct}, and Fr_{ct}. Similarly, $E(dP_{ct})$ and $E(\mathrm{Fr}_{ct})$ are the time series means of dP_{ct} and Fr_{ct}. "Other terms" at the end of eq (7) include the covariance term between Fr_{ct} and dP_{ct} as well as higher order terms. The first term on the right-hand side of (7), $Var(dP_{ct})\,[E(\mathrm{Fr}_{ct})]^2$, represents the contribution of the variance of dP_{ct} to the variance of π_{ct} and therefore is referred to as the intensive margin (IM) in the macroeconomics literature on price adjustments, while the

second term on the right hand side of (7), $Var(\mathrm{Fr}_{ct})\left[E(dP_{ct})\right]^2$, represents the contribution of the variance of Fr_{ct} to the variance of π_{ct}, which is referred to as the extensive margin (EM).

We calculate the variance of month-on-month inflation both for the CPI source data and for the scanner data, and decompose each into the intensive and extensive margins. To conduct this exercise, we need to identify products by their product ID (JAN code) to allow us to compare the price of a particular product at a particular outlet in a particular month and the price of the same product at the same outlet in the previous month in order to tell whether a price adjustment occurred. While we have such information for all products in the scanner data, this is not the case for the CPI data. For example, we do not have such information for products in item categories such as rice, ham, milk, and salt, since the JSB does not collect JAN code information for products belonging to these item categories. This means that in the rest of this section, we confine our analysis to the 46 items for which JAN code information is available.[10]

The result of the variance decomposition is presented in Table 6. We conduct variance decomposition item by item for the 46 items and aggregate the results attaching equal weights to all items (upper half of the table) and attaching the weights used in the CPI (lower half of the table). The column labeled "CPI" presents the results when using the CPI source data, showing that the variance of monthly inflation is 0.00076, which corresponds to a standard deviation of 0.028, and that the contribution of the extensive margin is 0.00001, while the contribution of the intensive margin is 0.00071, meaning that about 93 percent of inflation volatility stems from the intensive margin. The almost negligible contribution of the extensive margin implies that inflation volatility does not stem from time series fluctuations in the fraction of products that experience price adjustments. Note that the sum of the intensive and extensive margins does not coincides with the variance of inflation since the "other terms" in eq (7) are non-zero. However, the contribution of the "other terms" is very small and can be safely ignored.

Turning to the variance decomposition of scanner data-based inflation, which is presented on the column labeled "POS (Point-Of-Sale)," the variance of monthly inflation is now 0.00170, corresponding to a standard deviation of 0.041, which is more than twice as large as that of CPI inflation. However, the contribution of the intensive margin is dominantly

[10]The list of the 46 items is as follows: 1042, 1051, 1071, 1321, 1333, 1602, 1621, 1633, 1641, 1642, 1643, 1652, 1654, 1655, 1656, 1714, 1721, 1732, 1761, 1784, 1871, 1911, 1921, 1922, 1931, 1941, 1951, 2003, 2021, 4401, 4412, 4431, 4441, 4442, 4451, 4461, 6101, 6141, 9124, 9611, 9621, 9622, 9623, 9631, 9641, 9661. See Table 4 for item names corresponding to the item codes shown above.

large, as in the case of CPI inflation, accounting for about 84 percent of inflation volatility, while the contribution of the extensive margin is again negligible. Finally, the column labeled "Difference" shows that the difference between the variance of CPI inflation and the variance of scanner data-based inflation is 0.00094 and that most of this is accounted for by the difference in the intensive margin between the two indexes, which is 0.00071. The lower half of Table 6 repeats the same exercise with CPI item weights used in the aggregation and shows that the basic results remain unchanged. Finally, the decomposition results for each items are depicted in Figure 6 and again show that the contribution of the intensive margin dominates for each item.

Next, we examine why the intensive margin differs this much between CPI inflation and scanner data-based inflation. To address this, we decompose the difference in the intensive margin as follows:

$$\begin{aligned} \mathrm{IM}^{POS} - \mathrm{IM}^{CPI} = & \underbrace{\left[Var(dP)^{POS} - Var(dP)^{CPI}\right]\left[E(Fr)^{POS}\right]^2}_{\text{Due to the difference in } Var(dP)} \\ & + \underbrace{\left[\left(E(Fr)^{POS}\right)^2 - \left(E(Fr)^{CPI}\right)^2\right] Var(dP)^{CPI}}_{\text{Due to the difference in } [E(Fr)]^2} \end{aligned} \tag{8}$$

where IM^{POS} and IM^{CPI} are the intensive margins for scanner data-based inflation and CPI inflation. As the above equation indicates, the difference in the intensive margin can be decomposed into two parts: the part due to the difference in $Var(dP)$ and the part due to the difference in $[E(Fr)]^2$. Similarly, the difference in the extensive margin can be decomposed as follows:

$$\begin{aligned} \mathrm{EM}^{POS} - \mathrm{EM}^{CPI} = & \underbrace{\left[Var(Fr)^{POS} - Var(Fr)^{CPI}\right]\left[E(dP)^{POS}\right]^2}_{\text{Due to the difference in } Var(Fr)} \\ & + \underbrace{\left[\left(E(dP)^{POS}\right)^2 - \left(E(dP)^{CPI}\right)^2\right] Var(Fr)^{CPI}}_{\text{Due to the difference in } [E(dP)]^2} \end{aligned} \tag{9}$$

The result of this exercise is presented in the far right column of Table 6. One potential reason for the larger intensive margin in scanner data-based inflation is that there are greater fluctuations in the average size of individual price changes in the scanner data over time. However, what we find here is that $Var(dP)$ in the scanner data is in fact smaller, so that $\left[Var(dP)^{POS} - Var(dP)^{CPI}\right]\left[E(Fr)^{POS}\right]^2$ takes a negative value. The table shows that what instead contributes more to the difference in the intensive margin is the difference in

the probability of price adjustments, which is represented by $\left(E(Fr)^{POS}\right)^2 - \left(E(Fr)^{CPI}\right)^2$. That is, prices in the scanner data are less sticky than those in the CPI data, and this difference in price rigidity leads to the difference in the intensive margin and consequently to the difference in inflation volatility. This result remains unchanged even when CPI item weights are used in the aggregation, which is shown in the lower half of Table 6.

3.3.2 Probability of no price adjustments

Why are prices in the scanner data less sticky than prices in the CPI source data? To address this question, let us start by recalling that the event of no price adjustment occurs under the following circumstances: (1) the outlet to be sampled is not replaced this month, (2) the product to be sampled is not replaced this month, and (3) the price of the product at the outlet this month is the same as it was last month. All three conditions need to be satisfied simultaneously. For example, if the product sampled last month is replaced by a new one this month, the quality adjusted price of the new product is not identical with the price of the old product except by accident. The three conditions can be stated as

$$\begin{aligned} E(\text{Fr}) &= 1 - \Pr(\text{No price change}, \text{No product repl.}, \text{No outlet repl.}) \\ &= 1 - \Pr(\text{No price change} \mid \text{No product repl.}, \text{No outlet repl.}) \\ &\qquad \times \Pr(\text{No product repl.} \mid \text{No outlet repl.}) \\ &\qquad \times \Pr(\text{No outlet repl.}) \end{aligned} \tag{10}$$

where $E(\text{Fr})$ is the time series mean of Fr, which already appeared in eqs (7), (8) and (9), and $\Pr(\text{No price change}, \text{No product repl.}, \text{No outlet repl.})$ is the probability of no price adjustments. As indicated by the second equality of eq (10), the probability of no price adjustments can be decomposed into three parts. The first component, $\Pr(\text{No price change} \mid \text{No product repl.}, \text{No outlet repl.})$, is the probability that no price change occurs this month given that neither outlet replacement nor product replacement occurs this month. The second component, $\Pr(\text{No product repl.} \mid \text{No outlet repl.})$, is the probabiity that no product replacement occurs conditional on that no outlet replacement occurs. Finally, the third component, $\Pr(\text{No outlet repl.})$, is the unconditional probability that no outlet replacement occurs. We know from Table 6 that $E(\text{Fr})^{POS} > E(\text{Fr})^{CPI}$, but which of the three components contributes most to this inequality?

Table 7 shows the four probabilities appearing in eq (10). We calculate the four probabilities item by item and then aggregate them across items with no weights (shown in the

upper half of the table) as well as with the CPI weights (shown in the lower half of the table). The column labeled "CPI" presents the result for the CPI source data, which shows that the probability that no outlet replacements occur in any particular month is 0.979, indicating that outlet replacement takes place only at a very low probability. Next, the probability of no product replacement given that no outlet replacement occurs is 0.959, indicating that product replacement is also a rare event. Finally, the probability of no price change this month given that neither outlet replacement nor product replacement occurs this month is 0.765, implying that the conditional probability of price adjustment, sometimes referred to as the "frequency of price changes" in the literature, is 0.235, which is slightly higher than the figures obtained in previous studies on price stickiness in Japan such as Higo and Saita (2007). Multiplying these three probabilities, it turns out that Pr(No price change, No product repl., No outlet repl.) equal 0.720.

Comparing the results for the CPI source data and the scanner data with $k_S = 5$ and $k_P = 10$, we do not see any substantial difference in Pr(No outlet repl.) or in Pr(No product repl. | No outlet repl.). This means that the fact that a lower price rigidity is observed for the scanner data does not stem from more frequent outlet replacements and/or more frequent product replacements in the scanner data.[11] However, there exists a substantial difference between the CPI and scanner data in terms of the probability of no price change given that neither outlet replacement nor product replacement occurs (i.e., Pr(No price change | No product repl., No outlet repl.)). Specifically, the conditional probability of no price change is 0.765 for the CPI source data, while it is 0.652 for the scanner data with $k_S = 5$ and $k_P = 10$. It is this difference in the conditional probability that creates the difference in Pr(No price change, No product repl., No outlet repl.) between the CPI and scanner data. The lower panel of Table 7 shows that this result remains unchanged even when CPI weights are used in the aggregation. Figure 7 presents the four probabilities for each item. The figure indicates that the difference in the conditional probability of no price adjustment (i.e., Pr(No price change | No product repl., No outlet repl.)) tends to be higher for daily necessities such as toothbrushes (9611), hairdressing (9631), face lotion (9661), and insecticide (4451) than food and beverage items.

[11] The column labeled "POS with $k_S = 1$, $k_P = 1$" shows that there exists a non-trivial difference between the CPI and scanner data in terms of Pr(No product repl. | No outlet repl.), although the difference in Pr(No outlet repl.) is not that large. This can be interpreted as reflecting that, when k_P in eq (2) is set to $k_P = 1$, product replacement in the scanner data is much more frequent than it is when k_P is set to $k_P = 5$.

3.3.3 How to make scanner data based inflation less volatile

To make a further comparison between the CPI and scanner based price index in terms of price rigidity, we present price change distributions for individual products in the upper panel of Figure 8. The horizontal axis of Figure 8 shows a price change for a product i from month $t-1$ to month t, which is denoted by Δp_{it} (i.e., $\Delta p_{it} \equiv \ln P_{it} - \ln P_{it-1}$), while the vertical axis represents a density associated with each bin on the horizontal axis. Note that price change observations are pooled for all products and for all months in the sample period (January 2010-July 2014)

The upper panel of Figure 8 shows that the density associated with $\Delta p_{it} = 0$ is 0.774 for the CPI data while it is 0.646 for the scanner data, indicating again that prices in the CPI data are sticker than prices in the scanner data. More importantly, the price change distribution for the CPI has a dent near $\Delta p = 0$. Specifically, the density associated with the range $-0.04 < \Delta p < 0.04$ is significantly lower than the densities associated with outside that range. This is in a sharp contrast with the distribution for scanner data, which does not have such a dent. The dent near $\Delta p = 0$ for the CPI distribution means that small-sized prices are less likely to occur for the CPI data than for the scanner data. It is well known that small price changes are unlikely to occur if it incurs cost for price setters to change prices (i.e., the presence of menu costs in adjusting prices). Our finding suggests that menu costs play a more important role in the CPI data.[12] Finally, the densities associated with outside the near zero range tend to be higher for the scanner data than for the CPI data.

Our aim here is to transform individual prices in the scanner data such that the price change distribution for transformed prices comes closer to the price change distribution for the CPI data, thereby making scanner data based inflation less volatile. To achieve this, we apply a filter to the time series of P_{it} in the scanner data to obtain a new series, which is denoted by $\hat{P}_{it}$. Specifically, the filter we apply is as follows.

$$\hat{P}_{it} = \begin{cases} P_{it} & \text{with a probability of } \Lambda \\ \hat{P}_{it-1} & \text{with a probability of } 1-\Lambda \end{cases} \tag{11}$$

That is, $\hat{P}_{it}$ coincides with P_{it} with a probability of Λ, but it remains unchanged from its value in the previous month with a probability of $1-\Lambda$. This implies that $\hat{P}_{it}$ is stickier than P_{it}. We also assume that probability Λ depends on the percentage deviaiton of $\hat{P}$ from P as

[12] Why does a dent near $\Delta p = 0$ exist for CPI data but not so for scanner data? Why do menu costs play a more important role in CPI data? These are important questions but beyond the scope of this paper.

follows.

$$\Lambda(x_{it}) = a\left[1 - \exp(-bx_{it}^2)\right] \tag{12}$$

where x_{it} is defined by $x_{it} \equiv \ln \hat{P}_{it-1} - \ln P_{it}$, a and b are parameters with a taking a value between 0 and 1 and b taking a positive value. As indicated by eq. (12), $\Lambda(x)$ takes a minimum value when $x = 0$, and it monotonically increases as x deviates from zero. This means that, if $\hat{P}_{it-1}$ is close to P_{it}, it is unlikely to occur that $\hat{P}_{it}$ differs from $\hat{P}_{it-1}$. An important implication of this is that, since the size of price change for $\hat{P}$, which is given by $\hat{P}_{it} - \hat{P}_{it-1}$, is small if $\hat{P}_{it-1}$ is close to P_{it}, small-sized price changes are less likely to occur for $\hat{P}$ than for P.

The result of this exercise is shown in Figure 9,[13] but the effects of the filter can be seen more clearly by comparing the price change distribution for the filtered data, which is shown on the lower panel of Figure 8, with the corresponding distribution for the original data, which is shown on the upper panel of Figure 8. We see that the densities associated with small price changes, i.e., $-0.04 < \Delta p < 0.04$, are now much smaller than before, indicating that small price changes are less likely to occur for the filtered data than for the original data. The densities associated with small price changes are now closer to the corresponding densities for the CPI, although there still remain non-negiligible differences, especially for small price increases (i.e., $0 < \Delta p < 0.04$). Next, we see that the densities associated with outside the small price change range are now significantly lower than before, almost coinciding with the densities for the CPI. Finally, the density associated with $\Delta p = 0$ is now 0.739, which is much higher than it is for the original data (0.646), and comparable to the corresponding probability for the CPI data (0.774).

Next, we conduct decomposition of inflation volatility into intensive and extensive margins as we did for the original data. The result is presented in Table 8, showing that the variance of monthly inflation is 0.00083 for an unweighted average across items and 0.00064 for a weighted average. Comparing with the result for the original data, which is presented in Table 6, we see that inflation variance for the filtered data is substantially lower than for the original data (0.00110 for an unweighted average and 0.00084 for a weighted average), and closer to inflation variance for the CPI data (0.00065 for an unweighted average and 0.00051 for a weighted average). Tables 6 and 8 also show that a lower variance for the filtered data is mainly due to a lower intensive margin for the filtered data, which in turn is due to a higher price rigidity for the filtered data.

[13] Parameters a and b in eq. (12) are set at $a = 0.8$ and $b = 0.5$.

4 Conclusion

The use of scanner data in constructing consumer price indexes (CPIs) has become widespread practice in recent years among national statistical offices in a number of countries. Given this background, we examined in this paper how precisely one can reproduce the CPI constructed based on price surveys using scanner data. Specifically, we closely followed the procedure adopted by the Statistics Bureau of Japan when we sample outlets, products, and prices from our scanner data and aggregate them to construct a scanner data-based price index.

We showed that the following holds the key to precise replication of the CPI. First, the scanner data-based index crucially depends on how often one replaces the products sampled. The scanner data index shows a substantial deviation from the actual CPI when one chooses a value for the parameter associated with product replacement such that replacement occurs frequently. However, the deviation becomes much smaller if one picks a parameter value such that product replacement occurs only infrequently.

Second, even when products are replaced only infrequently, the scanner data index differs significantly from the actual CPI in terms of volatility. The standard deviation of the scanner data-based monthly inflation rate is 1.54 percent, which is more than three times as large as that for actual CPI inflation. We decomposed the difference in volatility between the two indexes into various factors, showing that it mainly stems from the difference in price rigidity for individual products. That is, actual CPI inflation is less volatile since individual prices in the CPI data are stickier. We propose a filtering technique to make individual prices in the scanner data stickier, thereby making scanner data-based inflation less volatile.

References

[1] Ariga, Kenn, and Kenji Matsui (2003), "Mismeasurement of the CPI," in M. Blomström, J. Corbett, F. Hayashi, A. Kashyap (eds.), *Structural Impediments to Growth in Japan*, University of Chicago Press, 89-128.

[2] Broda, Christian, and David E. Weinstein (2007), "Defining Price Stability in Japan: A View from America," *Monetary and Economic Studies*, Special Edition, Bank of Japan, December 2007, 169-189.

[3] Bureau of Labor Statistics (2007), "The Consumer Price Index," Chapter 17 in *BLS Handbook of Methods*. Available at `http://www.bls.gov/opub/hom/homch17.htm`.

[4] Chahrour, Ryan A. (2011), "Sales and Price Spikes in Retail Scanner Data," Economics Letters, Volume 110, Issue 2, February 2011, Pages 143-146.

[5] De Haan, Jan, Eddy Opperdoes, and Cecile Schut (1999), "Item Selection in the Consumer Price Index: Cut-off versus Probability Sampling," *Survey Methodology*, Vol. 25, No. 1, 31-41.

[6] Fenwick, David, Adrian Ball, Peter Morgan, and Mick Silver (2003), "Price Collection and Quality Assurance of Item Sampling in the Retail Prices Index: How Can Scanner Data Help?" in R. C. Feenstra and M. D. Shapiro (eds.), *Scanner Data and Price Indexes*, University of Chicago Press, 67-87.

[7] Fenwick, David (2014), "Exploiting new technologies and new data sources: the opportunities and challenges associated with scanner data," Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, May 26-28, 2014, Geneva.

[8] Handbury, Jessie, Tsutomu Watanabe, and David E. Weinstein (2013), "How Much Do Official Price Indexes Tell Us about Inflation?" NBER Working Paper No.19504, October 2013.

[9] Hautsch, Nikolaus (2004), *Modelling Irregularly Spaced Financial Data: Theory and Practice of Dynamic Duration Models*. Springer Science & Business Media.

[10] Higo, Masahiro, and Yumi Saita (2007), "Price Setting in Japan: Evidence from CPI Micro Data," Bank of Japan Working Paper Series, No.07-E-20, August 2007.

[11] Imai, Satoshi, and Tsutomu Watanabe (2014), "Product Downsizing and Hidden Price Increases: Evidence from Japan's Deflationary Period," Asian Economic Policy Review, Volume 9, Issue 1, 2014, 69-89.

[12] Imai, Satoshi, Chihiro Shimizu, and Tsutomu Watanabe (2012), "How Fast are Prices in Japan Falling?" CARF Working Paper Series, CARF-F-298, October 2012.

[13] International Labour Office (2004), Consumer Price Index Manual: Theory and Practice. Published for ILO, IMF, OECD, UN, Eurostat, The World Bank by ILO, Geneva.

[14] Shiratsuka, Shigenori (1999a), "Measurement Errors and Quality-Adjustment Methodology: Lessons from the Japanese CPI," Economic Perspectives 23(2), Federal Reserve Bank of Chicago, 2-13.

[15] Shiratsuka, Shigenori (1999b), "Measurement Errors in the Japanese Consumer Price Index," Monetary and Economic Studies 17(3), Institute for Monetary and Economic Studies, Bank of Japan, 69-102.

[16] Shiratsuka, Shigenori (2006), "Measurement Errors in the Japanese CPI," IFC Bulletin No.24, Irving Fisher Committee on Central-Bank Statistics, 36-43.

[17] Statistics Bureau of Japan (2012), "2010-Base Explanation of the Consumer Price Index," available at `http://www.stat.go.jp/english/data/cpi/1586.htm`

[18] Statistics Bureau of Japan (2013), "Outline of the Retail Price Survey," available at `http://www.stat.go.jp/english/data/kouri/pdf/outline.pdf`

[19] Watanabe, Kota, and Tsutomu Watanabe (2014), "Estimating Daily Inflation Using Scanner Data: A Progress Report" CARF Working Paper Series, CARF-F-342, February 2014.

Table 1: Number of Products that Meet the JSB Product Type Specifications

Item code	Description	No. of JAN codes (A)	No. of JAN codes that meet the product specifications (B)	(B/A)	Sales share of products that meet the product specifications
1001	Rice-A (domestic)	11962	1649	0.138	0.179
1002	Rice-B (domestic)	11962	1905	0.159	0.178
1011	Glutinous rice	477	321	0.673	0.935
1031	Boiled noodles	4944	1213	0.245	0.456
1041	Dried noodles	2194	37	0.017	0.002
1042	Spaghetti	1410	237	0.168	0.277
1051	Instant noodles	6879	6	0.001	0.063
1052	Uncooked Chinese noodles	8042	2439	0.303	0.268
1071	Wheat flour	199	71	0.357	0.597
1081	Mochi (rice cakes)	1687	1296	0.768	0.895
1151	Agekamaboko	20029	5129	0.256	0.291
1152	Chikuwa	3556	311	0.087	0.035
1153	Kamaboko	5917	4925	0.832	0.843
1161	Dried bonito fillets	897	9	0.010	0.001
1163	Shiokara (salted fish guts)	1870	989	0.529	0.645
1166	Fish prepared in soy sauce	1236	364	0.294	0.345
1173	Canned fish	1022	108	0.106	0.358
1252	Ham	2245	2065	0.920	0.973
1261	Sausages	5351	4753	0.888	0.940
1271	Bacon	2189	1936	0.884	0.906
1303	Milk	2144	1337	0.624	0.832
1311	Powdered milk	453	3	0.007	0.008
1321	Butter	369	30	0.081	0.458
1331	Cheese	599	23	0.038	0.242
1332	Cheese, imported	442	110	0.249	0.029
1333	Yogurt	557	174	0.312	0.610
1451	Azuki (red beans)	504	243	0.482	0.638
1453	Shiitake mushrooms	3700	57	0.015	0.006
1463	Dried tangle	980	536	0.547	0.482
1471	Bean curd	2914	2581	0.886	0.868
1472	Fried bean curd	2762	181	0.066	0.025
1473	Natto (fermented soybeans)	3809	3271	0.859	0.908
1481	Konnyaku (devil's tongue)	2705	2088	0.772	0.813
1482	Umeboshi, pickled plums	6743	5338	0.792	0.829
1483	Pickled radishes	4544	1383	0.304	0.317
1485	Tangle prepared in soy sauce	5339	2375	0.445	0.806
1486	Pickled Chinese cabbage	2818	1760	0.625	0.694
1487	Kimchi	5155	807	0.157	0.197
1491	Canned sweet corn	643	21	0.033	0.106
1591	Canned fruits	579	83	0.143	0.227
1601	Edible oil	1022	142	0.139	0.567
1602	Margarine	416	12	0.029	0.268
1611	Salt	1005	1	0.001	0.135
1621	Soy sauce	1793	24	0.013	0.234
1631	Soybean paste	5042	530	0.105	0.303
1632	Sugar	197	29	0.147	0.638
1633	Vinegar	636	2	0.003	0.222
1642	Ketchup	397	8	0.020	0.552

Item code	Description	A	B	B/A	Sales share of products that meet the product specifications
1643	Mayonnaise	451	3	0.007	0.205
1644	Jam	3823	5	0.001	0.081
1652	Instant curry mix	743	34	0.046	0.260
1653	Instant soup	1658	7	0.004	0.063
1654	Flavor seasonings	796	2	0.003	0.131
1655	Liquid seasonings	1758	9	0.005	0.339
1656	Granular flavor seasonings	776	2	0.003	0.000
1701	Yokan (sweet bean jelly)	3444	16	0.005	0.006
1711	Castella (sponge cakes)	2185	174	0.080	0.057
1714	Pudding	5280	4	0.001	0.171
1721	Biscuits	13130	4	0.000	0.021
1732	Candies	2067	22	0.011	0.162
1741	Sembei (Japanese crackers)	8314	453	0.054	0.035
1761	Chocolate	1238	8	0.006	0.257
1772	Peanuts	3651	705	0.193	0.124
1781	Chewing gum	1185	18	0.015	0.083
1782	Ice cream	1494	1	0.001	0.125
1791	Box lunch	21254	905	0.043	0.021
1793	Rice balls	7647	467	0.061	0.145
1794	Frozen pilaf	999	36	0.036	0.163
1811	Salad	11165	513	0.046	0.069
1812	Boiled beans	808	639	0.791	0.883
1851	Frozen croquettes	1167	64	0.055	0.039
1871	Cooked curry	3321	18	0.005	0.316
1881	Gyoza	3201	626	0.196	0.196
1891	Mazegohan no moto	303	3	0.010	0.367
1902	Green tea	5614	4329	0.771	0.602
1911	Black tea	1469	8	0.005	0.211
1914	Tea beverages	505	48	0.095	0.379
1921	Instant coffee	975	27	0.028	0.162
1922	Coffee beans	678	16	0.024	0.148
1923	Coffee beverages	3576	1184	0.331	0.620
1930	Fruit juice	2689	185	0.069	0.162
1931	Beverages which contain juice	2202	17	0.008	0.210
1941	Vegetable juice	353	2	0.006	0.307
1951	Carbonated beverages	400	4	0.010	0.047
1971	Fermented lactic drinks, sterilized (“Calpis”)	231	3	0.013	0.657
1981	Sports soft drinks	341	15	0.044	0.311
1982	Mineral water	1887	14	0.007	0.233
2003	Sake	6747	168	0.025	0.372
2011	Shochu (distilled spirits)	6691	32	0.005	0.172
2021	Beer	2430	246	0.101	0.391
2026	Low-malt beer	1389	157	0.113	0.308
2033	Whisky	1689	8	0.005	0.169
2041	Wine	21123	249	0.012	0.092
4401	Food wrap	993	14	0.014	0.180
4412	Facial tissue	1295	81	0.063	0.503
4413	Rolled toilet paper	2944	415	0.141	0.214
4431	Liquid detergent, kitchen	1212	21	0.017	0.076
4441	Detergent, laundry	866	144	0.166	0.457
4442	Fabric softener	836	43	0.051	0.410

Item code	Description	A	B	B/A	Sales share of products that meet the product specifications
4451	Insecticide	132	7	0.053	0.114
4461	Moth repellent for clothes	736	57	0.077	0.232
4471	Fragrance	1034	70	0.068	0.186
6095	Bath preparations	8648	54	0.006	0.059
6101	Sanitary napkins	2155	33	0.015	0.045
9111	Ball-point pens	15380	53	0.003	0.026
9115	Marking pens	1604	32	0.020	0.127
9121	Notebooks	13805	23	0.002	0.004
9124	Cellophane adhesive tape	1262	4	0.003	0.015
9127	Papers for office automation	518	97	0.187	0.766
9193	Dog food	2049	190	0.093	0.067
9195	Dry batteries	112	31	0.277	0.762
9196	Cat food	4250	580	0.136	0.332
9611	Toothbrushes	2388	32	0.013	0.102
9621	Toilet soap	2802	35	0.012	0.228
9622	Shampoo	4410	238	0.054	0.230
9623	Toothpaste	1255	21	0.017	0.110
9624	Hair conditioner	2932	138	0.047	0.185
9625	Hair dye	4200	37	0.009	0.077
9631	Hair liquid	380	2	0.005	0.255
9641	Hair tonic	233	5	0.021	0.192
9652	Face cream-B	1982	10	0.005	0.021
9661	Toilet lotion	5251	63	0.012	0.023
9672	Foundation-B	12600	74	0.006	0.024
9682	Lipsticks-B	18723	262	0.014	0.041
9692	Milky lotion-B	2157	18	0.008	0.018

Table 2: Number of Outlets, Products, and Observations in the Scanner Data

	No. of outlets	Entries	Exits	No. of products	No. of observations
2000	27	-	-	121,427	37,447,555
2001	27	0	0	129,848	40,632,653
2002	28	1	0	136,769	43,100,683
2003	28	0	0	136,663	39,347,502
2004	31	3	0	138,304	43,481,768
2005	28	1	4	135,222	44,197,393
2006	27	0	1	141,382	45,847,962
2007	32	5	0	146,165	44,291,942
2008	29	1	4	149,106	46,317,820
2009	28	0	1	142,518	45,808,810
2010	28	0	0	141,630	45,892,049
2011	28	0	0	143,821	45,559,906
2012	30	2	0	146,198	47,687,953
2013	32	3	1	151,387	50,038,122
2014*	30	0	2	124,933	29,430,411

* January 2014 to July 2014.

Table 3: Product Turnover in the 14 Outlets in Tokyo

	No. of products in the 14 outlets	Entries	Exits	Entry rate	Exit rate
2000	72,124	-	-	-	-
2001	73,961	26,224	24,387	0.355	0.330
2002	79,120	30,563	25,404	0.386	0.321
2003	80,656	28,660	27,124	0.355	0.336
2004	80,268	28,179	28,567	0.351	0.356
2005	81,887	29,176	27,557	0.356	0.337
2006	86,513	32,622	27,996	0.377	0.324
2007	91,633	33,822	28,702	0.369	0.313
2008	96,939	38,825	33,519	0.401	0.346
2009	92,065	32,421	37,295	0.352	0.405
2010	90,952	31,954	33,067	0.351	0.364
2011	90,339	30,379	30,992	0.336	0.343
2012	89,153	30,659	31,845	0.344	0.357
2013	93,729	32,292	27,716	0.345	0.296
2014 *	78,580	18,448	33,597	0.235	0.428

* January 2014 to July 2014.

Table 4: Items and Weights in the CPI Data

Code	Item	Weight
1001	Non-glutinous rice (single ingredient, “Koshihikari”)	23
1002	Non-glutinous rice (single ingredient, excluding “Koshihikari”)	31
1011	Glutinous rice	3
1031	Boiled “Udon” (wheat noodles)	8
1041	Dried “Udon” (wheat noodles)	6
1042	Spaghetti	4
1051	Instant noodles	11
1052	Uncooked Chinese noodles	11
1071	Wheat flour	2
1081	Mochi, rice-cakes	9
1151	Satsumaage, fried fish-paste patties	6
1152	Chikuwa, baked fish-paste bars	4
1153	Kamaboko, steamed fish-paste cakes	7
1161	Dried bonito fillets	3
1163	Shiokara, salted fish guts	1
1166	Gyokai-tsukudani, fish boiled in soy sauce	3
1173	Canned tuna fish	6
1252	Ham	16
1271	Bacon	6
1303	Fresh milk (sold in stores, in cartons)	36
1311	Powdered milk	2
1321	Butter	3
1333	Yogurt	23
1341	Hen eggs	18
1451	Azuki, red beans	2
1453	Shiitake, Japanese mushrooms, dried	2
1463	Dried tangle	3
1471	Tofu, bean curd	14
1472	Fried bean curd	8
1473	Natto, fermented soybeans	8
1481	Konnyaku, devil’s-tongue jelly	5
1482	Umeboshi, pickled Japanes apricot	9
1483	Pickled radishes	5
1485	Tangle prepared in soy sauce	4
1486	Pickled chinese cabbage	4
1601	Edible oil	7
1602	Margarine	2
1611	Salt	2
1621	Soy sauce	5
1631	Soybean paste	7
1632	Sugar	3
1633	Vinegar	4
1641	Worcester sauce	2
1642	Tomato ketchup	2
1643	Mayonnaise	4
1652	Instant curry mix	5
1654	Flavor seasonings	6

Code	Item	Weight
1655	Liquid seasonings	12
1656	Furikake, granular flavor seasonings	4
1701	Yokan, sweet bean jelly	14
1714	Pudding	7
1721	Biscuits	13
1732	Candies	7
1741	Senbei, Japanese rice crackers	19
1761	Chocolate	17
1772	Peanuts	2
1783	Potato chips	11
1784	Jelly	9
1791	Box lunch	47
1811	Salad	18
1851	Frozen croquettes	9
1871	Cooked curry	3
1881	Gyoza, Chinese meat dumpling	15
1902	Green tea (“Sencha”)	16
1911	Black tea	4
1921	Instant coffee	6
1922	Coffee beans	6
1931	Fruit drinks (20-50% fruit juice)	6
1941	Vegetable juice	10
1951	Cola drinks	12
2003	Sake	14
2021	Beer	33
2033	Whisky (40% or more and less than 41% alcohol)	3
2041	Wine	4
4401	Food wrap	3
4412	Facial tissue	7
4413	Rolled toilet paper	10
4431	Liquid detergent, kitchen	9
4441	Detergent, laundry	11
4442	Fabric softener	4
4451	Insecticide	6
4461	Moth repellent for clothes	1
6101	Sanitary napkins	9
6141	Disposable diapers (baby)	5
9121	Notebooks	6
9124	Cellophane adhesive tape	2
9195	Dry batteries	5
9611	Toothbrushes	3
9621	Toilet soap	3
9622	Shampoo	9
9623	Toothpaste	6
9631	Hair dressing	8
9641	Hair tonic	4
9661	Face lotion	20

Table 5: Root Mean Squared Error for the Discrepancy between Scanner Data-Based Inflation and Actual CPI Inflation

	$k_S = 1$	$k_S = 2$	$k_S = 3$	$k_S = 4$	$k_S = 5$
$k_P = 1$	0.0105	0.0103	0.0103	0.0102	0.0102
$k_P = 2$	0.0095	0.0095	0.0094	0.0089	0.0094
$k_P = 3$	0.0090	0.0089	0.0089	0.0088	0.0088
$k_P = 4$	0.0087	0.0089	0.0086	0.0085	0.0085
$k_P = 5$	0.0085	0.0083	0.0084	0.0084	0.0083
$k_P = 6$	0.0084	0.0082	0.0082	0.0082	0.0082
$k_P = 7$	0.0082	0.0080	0.0081	0.0081	0.0081
$k_P = 8$	0.0082	0.0080	0.0080	0.0080	0.0079
$k_P = 9$	0.0082	0.0080	0.0080	0.0080	0.0080
$k_P = 10$	0.0081	0.0080	0.0078	0.0079	0.0078

Table 6: Decomposition of Inflation Volatility into Extensive and Intensive Margins

Unweighted Average Across Items			
	CPI (A)	POS (B)	Difference (B)-(A)
$Var(\pi_{ct})$	0.00065	0.00110	0.00045
Extensive Margin	0.00001	0.00001	0.00000
Due to $Var(\mathrm{Fr}_{ct})$			0.00000
Due to $[E(dP_{ct})]^2$			0.00000
Intensive Margin	0.00063	0.00132	0.00068
Due to $Var(dP_{ct})$			-0.00032
Due to $[E(\mathrm{Fr}_{ct})]^2$			0.00100
Weighted Average Across Items			
	CPI (A)	POS (B)	Difference (B)-(A)
$Var(\pi_{ct})$	0.00051	0.00084	0.00032
Extensive Margin	0.00000	0.00001	0.00001
Due to $Var(\mathrm{Fr}_{ct})$			0.00000
Due to $[E(dP_{ct})]^2$			0.00000
Intensive Margin	0.00048	0.00097	0.00049
Due to $Var(dP_{ct})$			-0.00028
Due to $[E(\mathrm{Fr}_{ct})]^2$			0.00078

Table 7: Probability of No Price Adjustment

Unweighted Average Across Items			
	CPI	POS $k_S = 5$ $k_P = 10$	POS $k_S = 1$ $k_P = 1$
Pr(No outlet repl.)	0.979	0.992	0.983
Pr(No product repl. | No outlet repl.)	0.959	0.956	0.898
Pr(No price change | No product repl., No outlet repl.)	0.765	0.652	0.656
Pr(No price change, No product repl., No outlet repl.)	0.720	0.619	0.580
Weighted Average Across Items			
	CPI	POS $k_S = 5$ $k_P = 10$	POS $k_S = 1$ $k_P = 1$
Pr(No outlet repl.)	0.980	0.993	0.986
Pr(No product repl. | No outlet repl.)	0.960	0.946	0.878
Pr(No price change | No product repl., No outlet repl.)	0.772	0.645	0.650
Pr(No price change, No product repl., No outlet repl.)	0.728	0.607	0.564

Table 8: Inflation Volatility for Filtered Price Data

Unweighted Average Across Items			
	CPI (A)	POS (B)	Difference (B)-(A)
$Var(\pi_{ct})$	0.00065	0.00083	0.00019
Extensive Margin	0.00001	0.00001	0.00001
Due to $Var(\mathrm{Fr}_{ct})$			0.00000
Due to $[E(dP_{ct})]^2$			0.00000
Intensive Margin	0.00063	0.00095	0.00031
Due to $Var(dP_{ct})$			-0.00001
Due to $[E(\mathrm{Fr}_{ct})]^2$			0.00030
Weighted Average Across Items			
	CPI (A)	POS (B)	Difference (B)-(A)
$Var(\pi_{ct})$	0.00051	0.00064	0.00013
Extensive Margin	0.00000	0.00001	0.00001
Due to $Var(\mathrm{Fr}_{ct})$			0.00000
Due to $[E(dP_{ct})]^2$			0.00000
Intensive Margin	0.00048	0.00070	0.00022
Due to $Var(dP_{ct})$			-0.00002
Due to $[E(\mathrm{Fr}_{ct})]^2$			0.00025

Figure 1: Replication Using the CPI Source Data

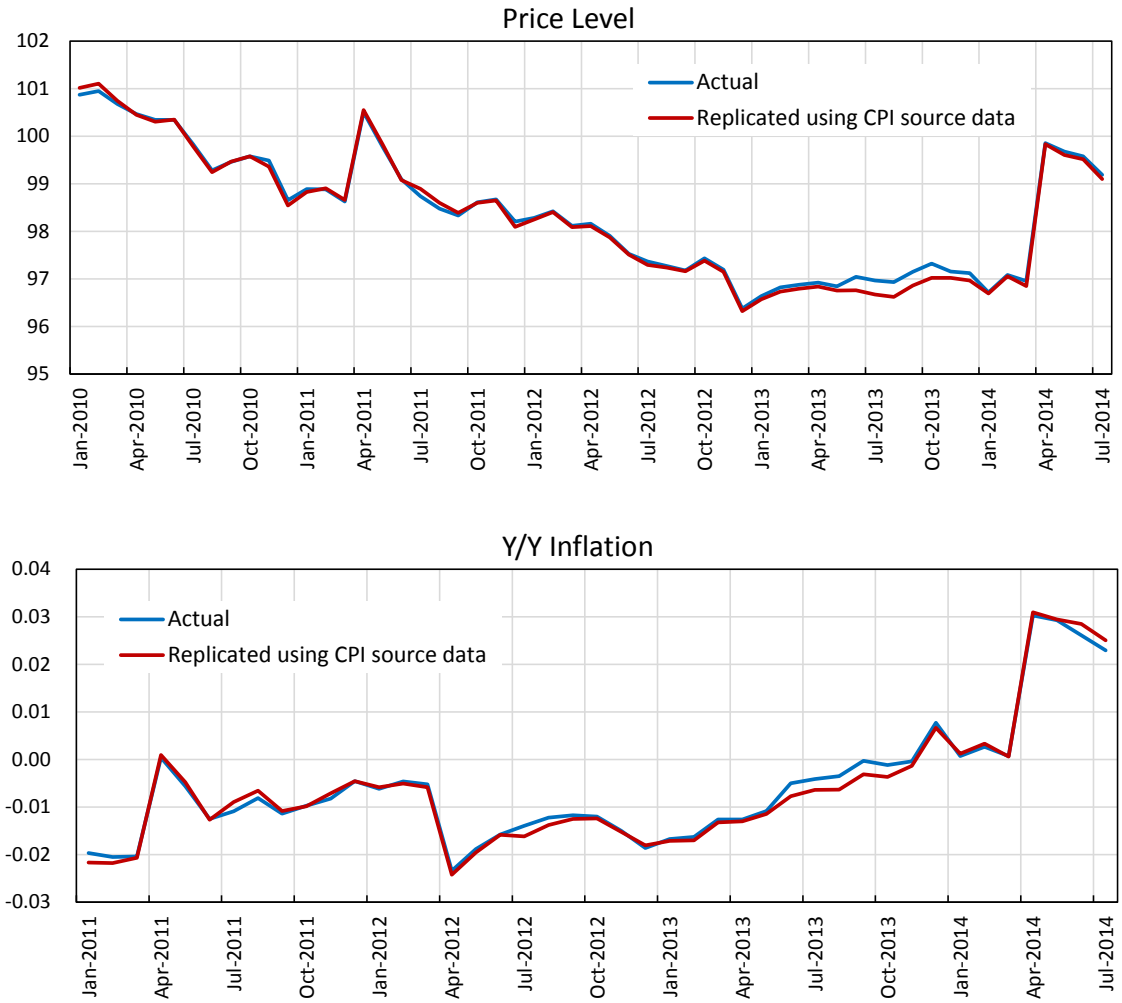


Figure 2: Replication Using the Scanner Data

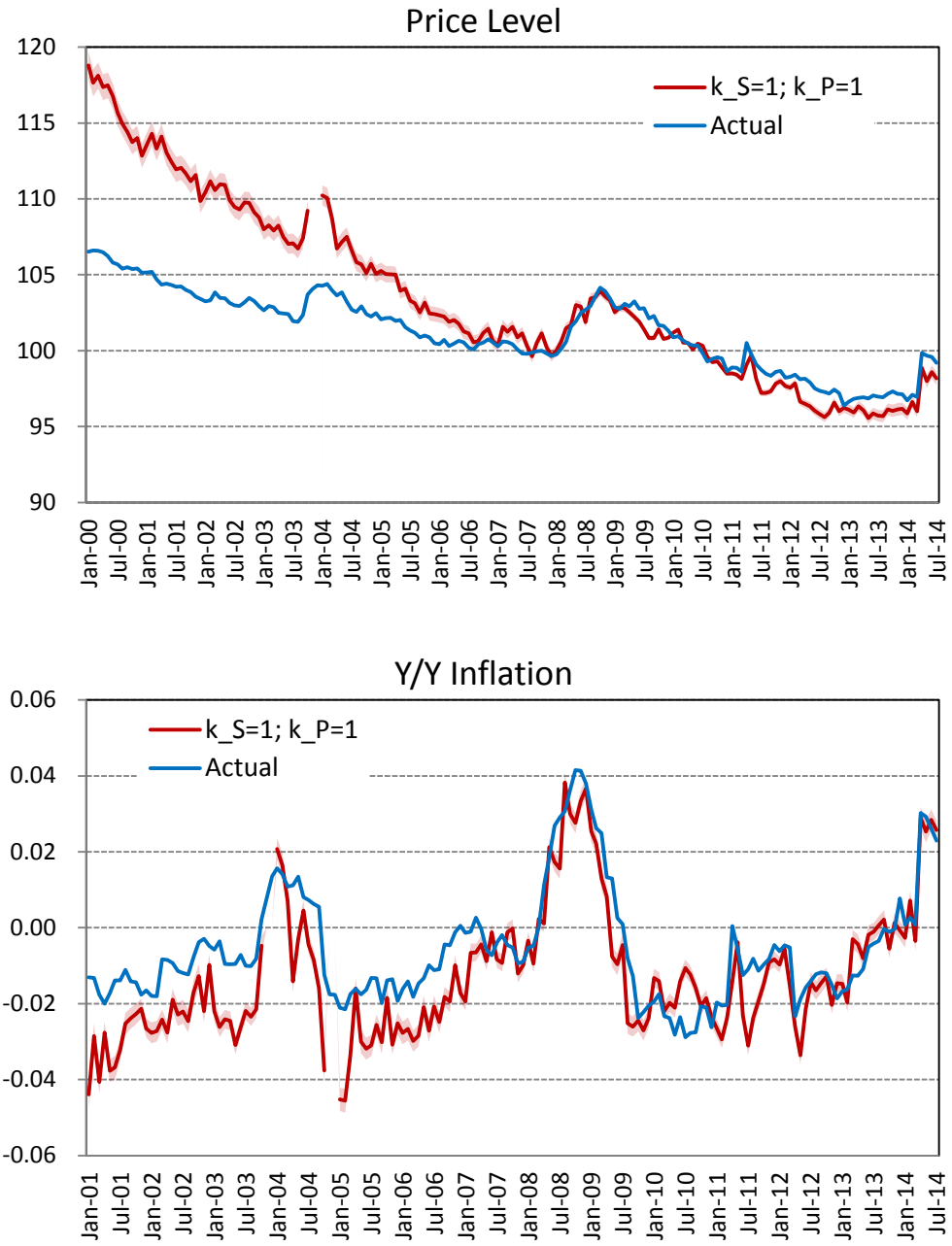


Figure 3: Root Mean Square Error for the Difference between Scanner-Data Based Inflation and CPI Inflation

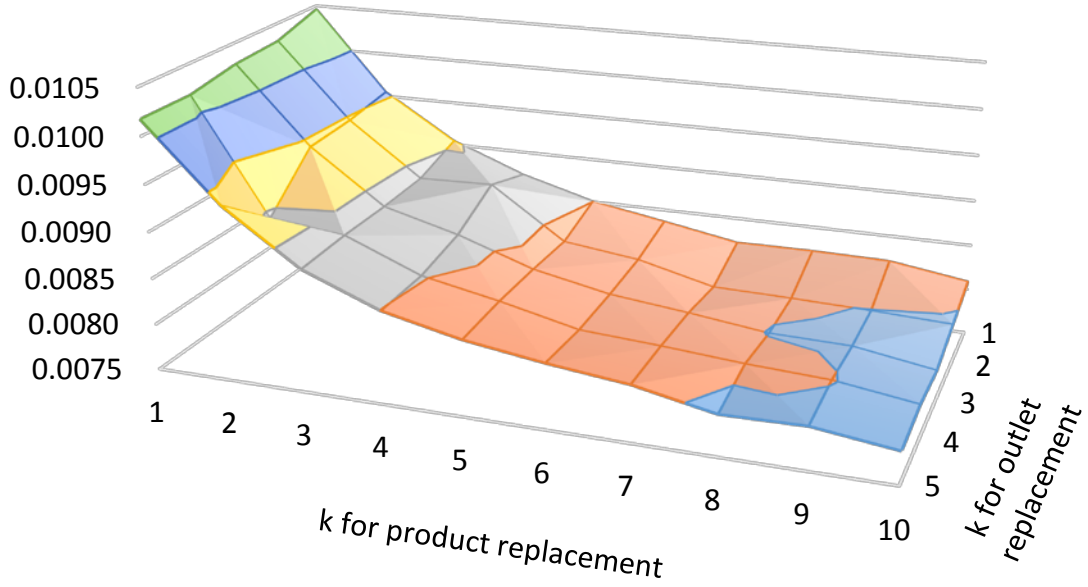


Figure 4: Inertia in Outlet and Product Replacement

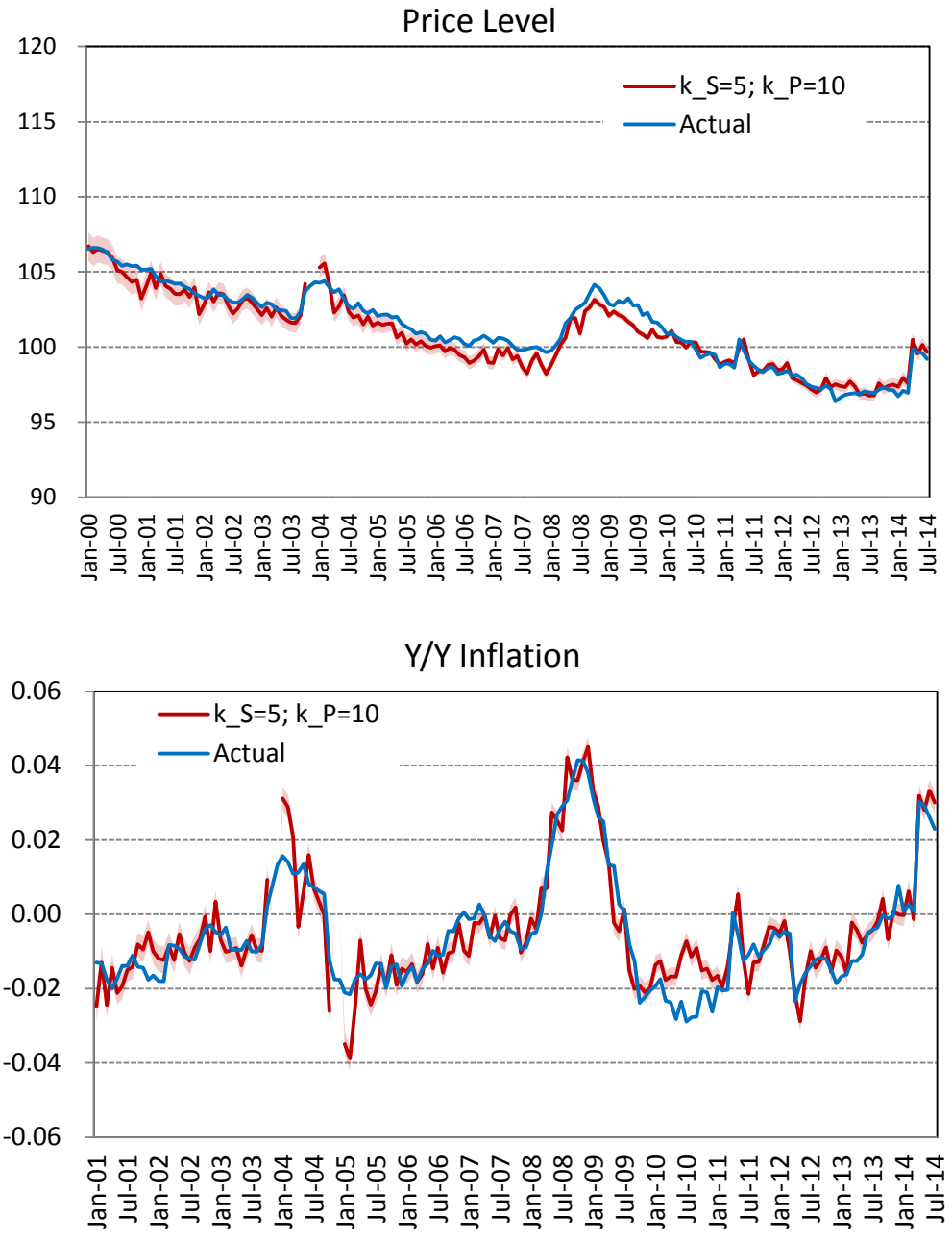


Figure 5: Mean and SD of Monthly Inflation By Item

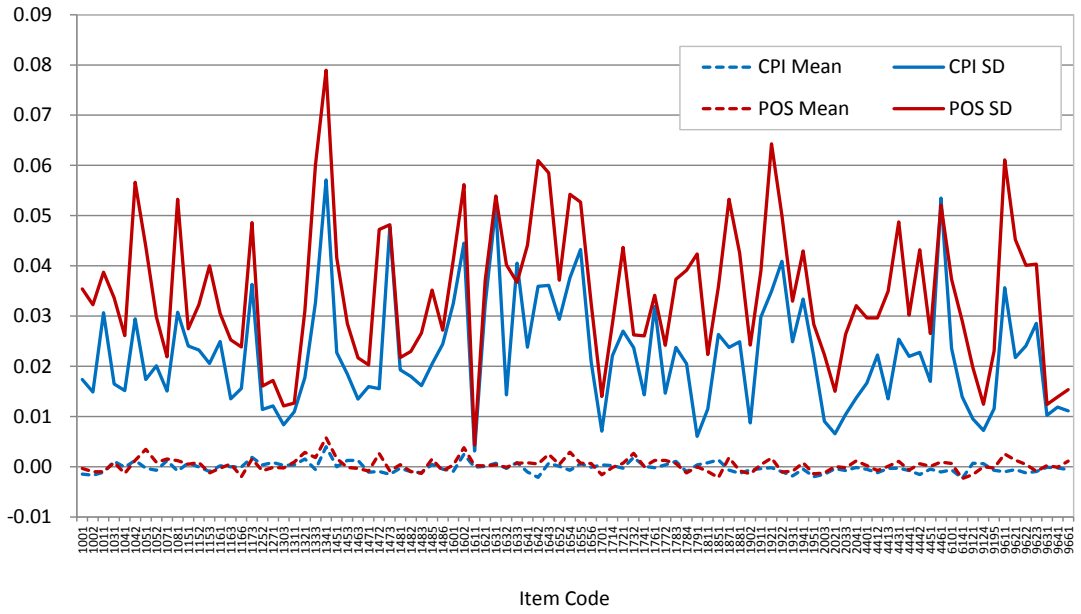
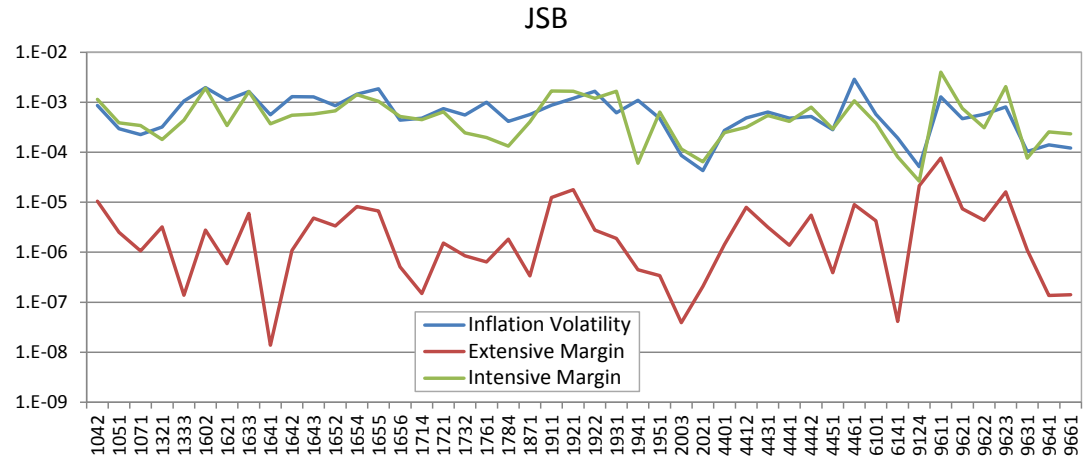


Figure 6: Decomposition of Inflation Volatility into Extensive and Intensive Margins



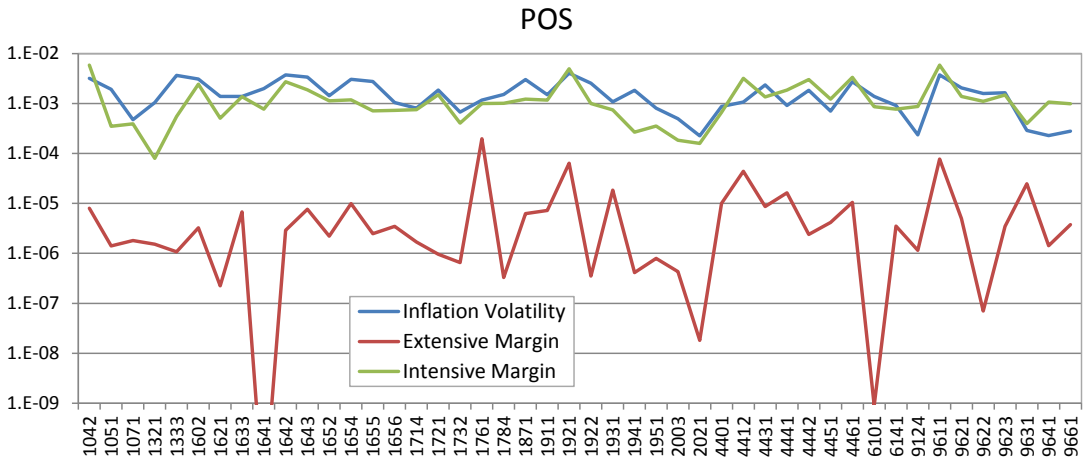


Figure 7: Probability of No Price Adjustment

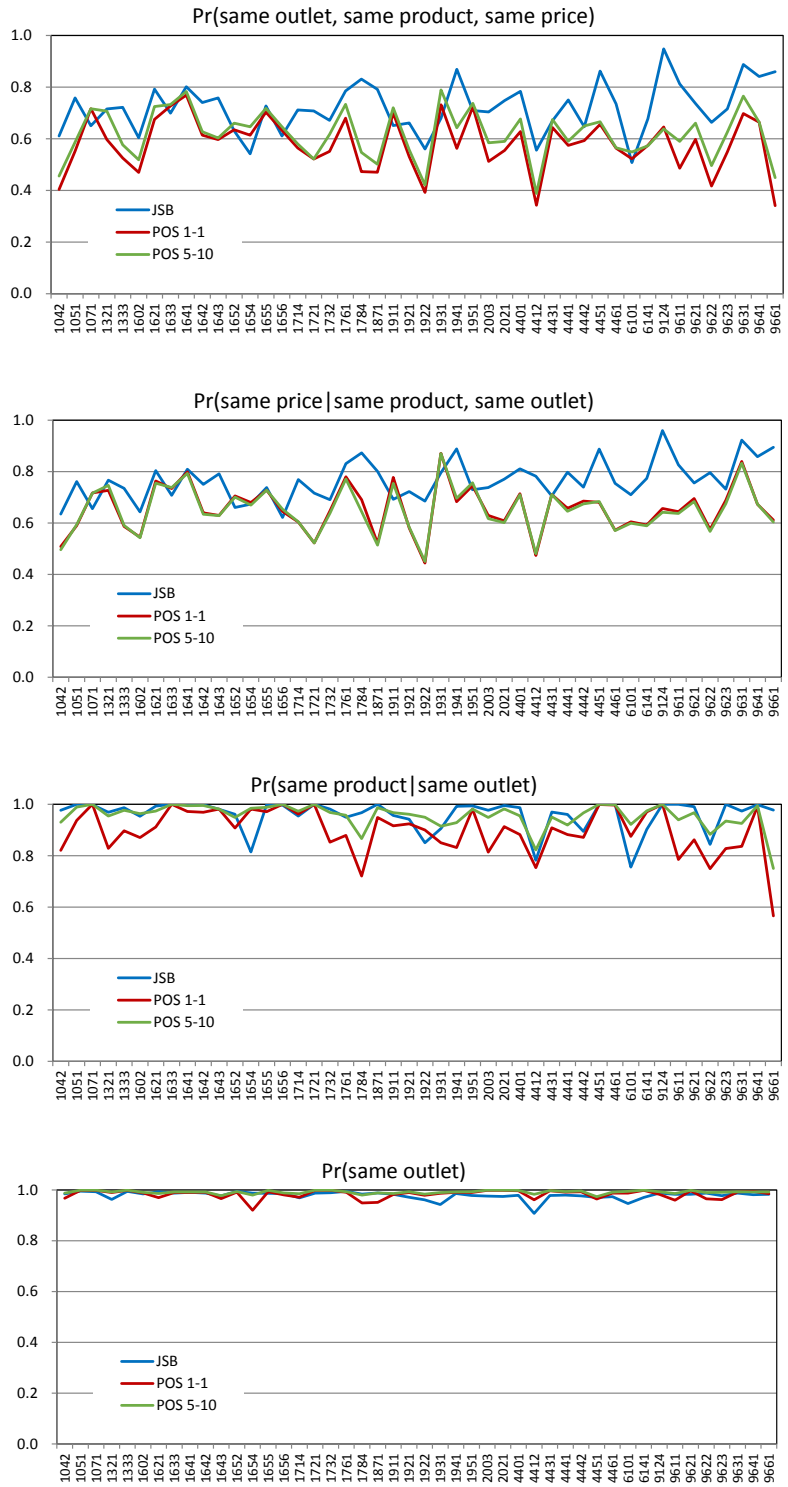


Figure 8: Price Change Distributions

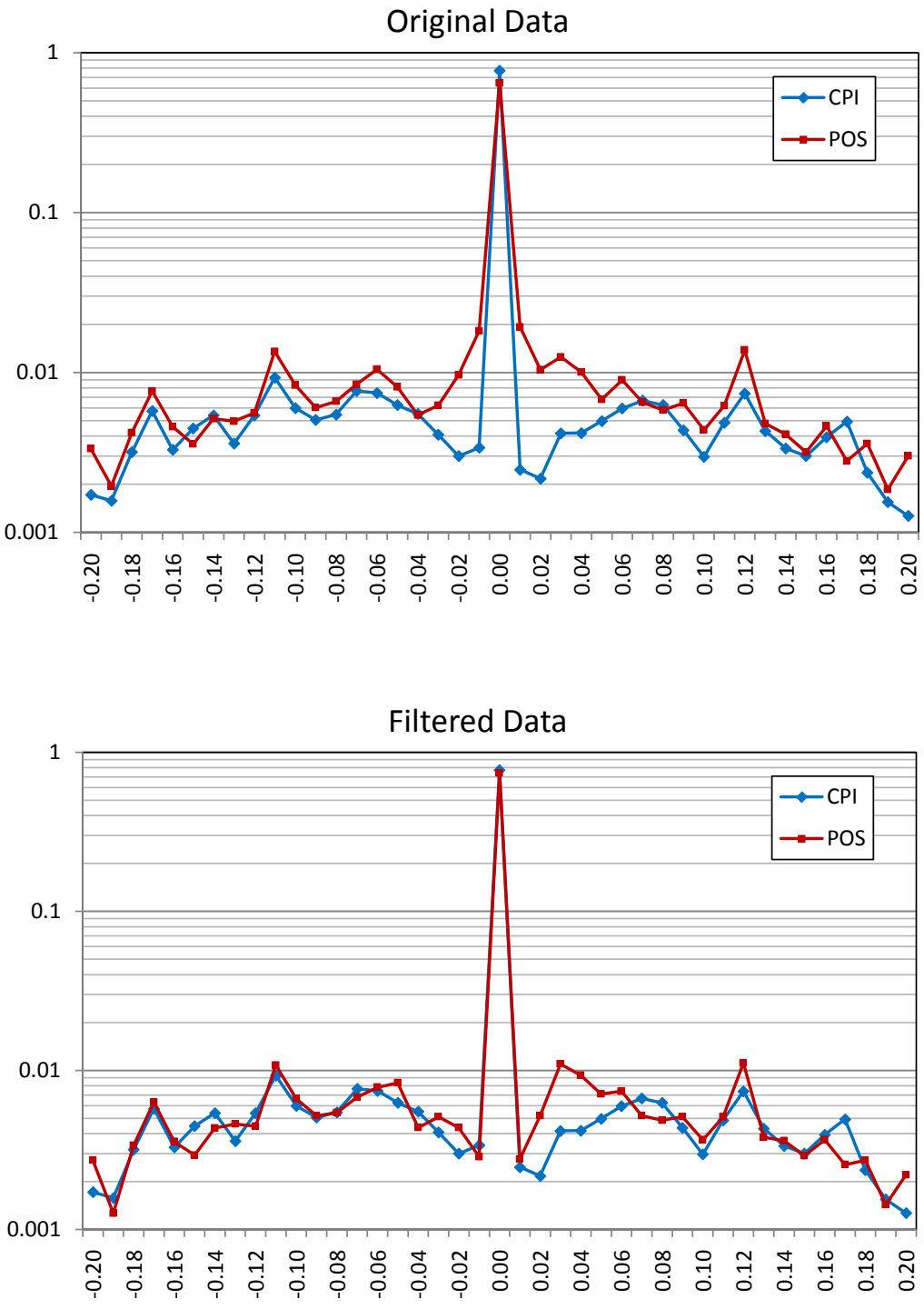


Figure 9: Replication Using Filtered Price Data