ESTIMATING MISSING PRICES IN PRODUCER PRICE INDEX

December 2006

Onimissi M. Sheidu
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 3650, Washington, DC 20212.
Sheidu.onimissi@bls.gov

The Producer Price Index (PPI) is a monthly estimate of average changes in prices received by domestic producers of goods and services in all stages of processing. Each month the PPI requests data for over 100,000 price quotes. For those data not received, the PPI must estimate a value to be used in the index. In this paper we investigated whether there is an added advantage in terms of the estimates by using only the weighted relatives of items in a cell with similar products to estimate an item's missing price, or to use higher aggregate cell relatives comprised of different product cells. We discovered that the proposed method of estimating missing prices using detailed product cell relatives is superior to using the aggregated product cell relatives only in cases where there are enough items, usually greater than 10, in a cell; otherwise, using aggregates of these product cell relatives produces more accurate estimates.

Key words: Producer Price Index, Weighted Relatives, Aggregate Cell Relatives, Missing Prices, Absolute difference, simulation.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

1. INTRODUCTION

The Producer Price Index (PPI) is a key economic indicator that measures the average change over time in selling prices received by domestic producers. It serves as one of the nation's inflationary indicators, particularly for the business sector of the economy.

The PPI publishes three primary outputs. These are industry output indexes, commodity indexes, and stage of processing indexes. The industry structure upon which the PPI is based is the North American Industry Classification System (NAICS). Under NAICS, establishments that use the same or similar processes to produce goods and services are grouped together. Each month the PPI publishes indexes at the 6-digit NAICS level along with more detailed product level indexes. In addition the PPI produces detailed commodity indexes. The data collected by industry are regrouped into commodity classifications without regard to the particular industry in which they are produced. These data are sometimes referred to as wherever made indexes. The PPI's stage of processing (SOP) indexes is aggregations of the commodity indexes. Theses indexes are used to measure inflation as it passes from one stage of the U.S. economy to another. The three main SOP indexes are Finished Goods, Intermediate Goods, and Crude Goods. Refer to the BLS Handbook of Methods (Chapter 14).

Indexes are calculated by moving the weights of each item by its monthly price change. The weights are based on sampling factors and revenue. The items are grouped into similar product categories, called product cells. The index relatives for these product cells are then weighted together into more aggregate cells.

If a company does not report a price for a particular month, its weight must be moved by an estimate of its price change. In the PPI the primary method of imputing for missing prices is called the cell relative method. This process assigns the average price change for all reported prices in the cell to the missing items. One benefit of this method is that the index value for the cell is the same, whether the items with missing prices are included or not.

Since 2003, product cells are assigned at the 8-digit Census (NAICS) code level – or at the 10-digit level under certain conditions. For the purpose of this research we distinguished between two types of cell indexes: the 7-digits cell indexes for the calculation of PPI indexes at the Census 7-digit NAICS code level and the 8-digits cell indexes (product relatives) for the calculation of PPI indexes at the Census 8-digit NAICS code level. It is helpful to note that the 8-digit NAICS code cell level is more detailed and homogeneous containing items of similar products than the 7-digit NAICS code cell level. Thus, the 7-digit cell index is an aggregation of one or more eight-digit cell indexes under the same NAICS. A further aggregation process, of course, yields an index of a particular NAICS, which is the 6-digit (NAICS).

The goal of this study is to determine the best imputation level for point estimate. We do not consider variance or the distribution of prices as a factor in deciding which method to use. Section 2, Overview and Methods of PPI Index Calculation presents an overview of the study and methods of index calculation used in this study. Section 3, Simulation describes the sampling procedures, simulation study and the computed statistics. Section 4, Results summarizes the result for overall cells, as well as taking the individual cell specifics into consideration. And finally, section 5 gives recommendation on the approach to calculating cell relatives with different level of items presentation.

2. OVERVIEW AND METHODS OF PPI INDEX CALCULATION

2.1 PPI Sampling Process

The source of the PPI sampling frame is the BLS Longitudinal Data Base (LDB), which contains U.S. business frame records representing all U.S. non-farm industries, with the exception of some sole proprietors. Data for the LDB are collected under the ES-202 program. This is a Federal/State Cooperative Program conducted with State Employment Security Agencies (SESAs). The LDB consists of all covered employers under the Unemployment Insurance (UI) Tax System, except for self employed and small family businesses. The street address and number of establishments comprising a company record are not priority data elements.

Although the UI-based LDB has been found to be adequate for use in sampling PPI industries, it has been preferable to use an alternative frame for some industries, especially in the service sector, when information from the alternative frame source has been deemed more accurate and complete than the UI data. When an alternative frame is used, there have been several size measures other than employment that have been used in sampling PPI industries such as Newspapers--circulation; Cable TV—subscribers and Hospitals-- admissions.

The Producer Price Indexes are based on probability samples. The selection of the items to be used in an industry index is done in two stages. First-stage sampling is the selection of establishments proportional to measure of size (employment) from a frame of all establishments that produce in a particular industry. The frame unit's shipments and receipts would be the desired measure of size, but these are not always readily available in the frame. Employment is used since it is available and is highly correlated with shipments and receipts. A unit may be so large that its probability of being sampled is 100%. Any such units are called certainty units. In first-stage sampling, all certainty units are removed before the other units, called probability units, are selected. The final sample consists of certainty units and probability units.

Second-stage sampling results in the selection of the given number of items from the first-stage unit. The BLS Field Economist (FE), with the cooperation of the company respondent, performs the selection of the actual items for use in the PPI indexes. To improve the likelihood of publishing every cell, PPI developed a procedure to allow the FE to spread the items assigned for collection across item categories prior to actual item selection. The item categories correspond to predetermined publication cells and are preprinted on a worksheet called the Industry Specific Disaggregation Worksheet (ISDWS). The FE organizes the list of items provided by the respondent into these categories (called ISDWS categories) and enters the shipments and receipts for each category. The FE assigns items to categories by using a combination of certainty and PPS selection methods to guarantee full representation of all item lines produced by an establishment in the sample.

2.2 Index calculation

The goal of PPI is to provide accurate monthly average change in selling prices received by domestic producers of goods and services. This monthly average is made up of over 100,000 price quotes per month in the form of item prices from more than 30,000 sample establishments. Each price quote represents an individual item code, which are further grouped in the form of cells according to their product similarity. These items, which are usually initiated and re-priced every month, are used to compute indexes of price change.

PPI uses Laspeyres index of price change to calculate monthly index aggregation.

This generally involves the calculation of the cell or lowest level indexes using item weights from the sample and the calculation of the aggregate or higher level indexes using adjusted Census weights. A simple formula for a Laspeyres index is

$$\text{Laspeyres Index} = \frac{\sum P_c Q_b}{\sum P_b Q_b} \; x \;\; 100$$

Where,

Σ = The sum over all of the items included in the index

P_c = Current period price of a certain item

P_b = Base period price of the same item

Q_b = Base period quantity of the same item

When the above formula is written as a weighted average of price relatives, it gives us:

Laspeyres Index $= \frac{\sum P_c Q_b}{\sum P_b Q_b}\ x\ 100$ =

$$\frac{\sum\left(\frac{P_c}{P_b}\ x\ P_b Q_b\right)}{\sum P_b Q_b}\ x\ 100 = \frac{\sum w\ \frac{P_c}{P_b}}{\sum w}\ \text{x 100}$$

Where,

w = $P_b Q_b$ weight,

$\frac{P_c}{P_b}$ = long-term price relative for each individual item

As mentioned above an item is defined as a product, unique with respect to price determining characteristics. These items are selected by a systematic sampling method from the universe of products manufactured by a company. The values for the items used in calculation of the indexes are the product of the long term price relative of the item i and the weight of the item i. This can be expressed as the item relative, $\mathrm{I}_{ij,c}$ for an item i belonging to establishment j at time c, by

$$\mathrm{I}_{ij,c} = w_i\ x\ \frac{P_c}{P_b} = w_{ij} r_{ij,C}$$

Where,

$r_{ij,c}$ is the long term price relative from the base period b to time c for item j in establishment j,

$w_i = w_{ij}$ is the item weight which is derived from the probability of selection of establishment j, and the revenue of establishment j represented by item i at time b.

The index for each industry has a defined aggregation structure. This structure includes the detailed product cells to which items are assigned and the order of aggregation, or combination, of the lowest level cells to form higher-level aggregate cells.

In this study the indexes were computed only at the 7-digit and 8-digit NAICS cells levels. These indexes, which are long-term indexes representing change from the base period b to time c, were computed using the following equation:

$$CellIndex_C = \frac{CellAggregate_C}{\sum w_i} = \frac{\sum\left(w_i\ x\ \frac{P_c}{P_b}\right)}{\sum w_i} = \frac{\sum w_i r_c}{\sum w_i}\ x\ 100$$

Where,

$\sum =$ the sum over all items in the cell

C = the current month

$r_c = \frac{P_c}{P_b}$ = long-term price relative for each individual item at time c.

The problem with using this method directly is that the sample of items in the index is, in reality, constantly changing. Items may be added to or deleted from a cell as the result of the addition or discontinuation of respondents. In addition, the estimation of missing prices may require weight to be implicitly moved at a higher level, and the effect on the cell is the same as if the weight had actually been removed from the cell in that month. As a result, a simple comparison of the sum of the current weight to the base period weight would compare two different samples and, therefore, reflect sample changes in addition to price changes.

For the above reasons indexes are calculated using a modified Laspeyres index formula. The modification differs from the conventional Laspeyres in that the PPI uses a chained index instead of a fixed-base index. Chaining involves multiplying an index (or long term relative) by a short term relative (STR). This is useful since the product mix available for calculating indexes of price change can change over time. These two methods produce identical results as long as the market basket of items does not change over time and each item provides a usable price in every period. In fact, due to non-response, the mix of items used in the index from one period to the next is often different. The benefits of chaining over a fixed base index include a better reflection of changing economic conditions, technological progress, and spending patterns, all of which could lead to item substitution in some establishments, and a suitable means for handling items of nonresponse or refusal for certain period of time.

Below is the derivation of the modified fixed quantity Laspeyres formula used in the PPI.

$$LTR_c = \left(\frac{\sum p_{i,c} q_{i,b}}{\sum p_{i,b} q_{i,b}} \right) (100) = \left(\frac{\sum \frac{p_{i,c}}{p_{i,b}} p_{i,b} q_{i,b}}{\sum p_{i,b} q_{i,b}} \right) (100)$$

$$LTR_c = \left(\frac{\sum w_{i,b} r_{i,c}}{\sum w_{i,b}} \right) (100) = \left(\frac{\sum w_{i,b} r_{i,c}}{\sum w_{i,b} r_{i,c-1}} \right) \left(\frac{\sum w_{i,b} r_{i,c-1}}{\sum w_{ij,b}} \right) (100)$$

$$LTR_c = \left(\frac{\sum w_{i,b} r_{i,c}}{\sum w_{i,b} r_{i,c-1}} \right) \left(\frac{\sum w_{i,b} r_{i,c-1}}{\sum w_{ij,b}} \right) = (STR_c)(LTR_{c-1})$$

Where:
LTR c = long term relative of a collection of items at time c,
p_{ic} = price of item *i* at time *c*,
$q_{i,b}$ = quantity of item *i* in base period b,
$w_{i,b} = (p_{i,b})(q_{i,b})$ = total revenue of item *i*, in base period b,
$r_{i,c} = p_{i,c} / p_{i,b}$ = long term relative of item *i* at time *c*,

$STR_C = \dfrac{\sum w_{i,b} r_{i,c}}{\sum w_{i,b} r_{i,c-1}}$ = Short-term relative of a collection of items *i*, at time *c*

In this formula as we can see, each term is only a price comparison because the sample of items used to calculate a particular term, which is constant. It is always the sample available in the month represented by the numerator that is to be compared in each term. These changes, which represent only price change, are then multiplied (or linked) together to obtain the actual change in the index level over the specified period. In addition to possible changes in the structure of the cell, as discussed previously, the PPI has a four-month price correction policy. This policy allows prices previously used in index calculation to be revised for a period of up to four months after the initial calculation of the index. In order to reflect all revisions that may have occurred, the aggregates for the previous four months are recalculated each month and the current month's cell index is then calculated by chaining back to the last month in which a change may have occurred. Thus in light of the above, the actual calculation formula is expressed as:

$$CellIndex_C = \left(\frac{\sum w_{i,b} r_{i,c}}{\sum w_{i,b} r_{i,c-1}} \right) \left(\frac{\sum w_{i,b} r_{i,c-1}}{\sum w_{i,b} r_{i,c-2}} \right) \left(\frac{\sum w_{i,b} r_{i,c-2}}{\sum w_{i,b} r_{i,c-3}} \right) \left(\frac{\sum w_{i,b} r_{i,c-3}}{\sum w_{i,b} r_{i,c-4}} \right) (Cell\ Index_{c-4})$$

Where,
$\sum w_{i,b} r_{i,c}$ = Cell Aggregate at time c
C = the current month
C-1, C-2, C-3, C-4 = months 1, 2, 3, and 4 prior to the current C month, C-4 is the last month allowed for index revision).

Since no changes in structure or revisions in prices are allowed beyond C-4, the cell index for C-4 is calculated as:

$$CellIndex_{C-4} = \left(\frac{\sum w_{i,b} r_{i,c-4}}{\sum w_{i,b} r_{i,c-5}} \right) Cell\ Index_{c-5})$$

Because there are no changes allowed beyond C-4, it is never necessary to chain back beyond this point.

Since items prices are collected over time, (from month to month), it becomes imperative to estimate or impute the item price quotes at a given month that are missing in order to accurately estimate the overall index aggregate.

PPI uses cell relative to estimate-missing prices. This is a process in which the mean value of the relative price changes for respondents (items with prices) is assigned to all nonrespondents (items with missing prices) within that particular cell. In other words, the cell relative method assigns to items with missing prices, a price change equal to the average price change for all items in their cell that actually have good prices.

3. SIMULATION STUDY

3. 1 Sample Selection:

We based the selection of cells for this study on the following criteria:

1. Cell must have multiple price changes, that is, all cells with only unchanged item prices are excluded from the sample.
2. Seven-digit cell level must contain multiple product codes (8-digit cell level), in other words, there must be at least two eight digit level cells in a particular 7 digit cell level to be considered fit for the current study.

Cells were judgmentally as well as randomly selected to obtain a diverse group of cells in order to adequately represent the population. Out of the 23 NAICS selected

(7digit cell level), nine cells initially were judgmentally chosen and 16 were randomly chosen. We started with a list of several cells that have multiple product codes. Of these, nine were found to have a sufficient number of items as well as multiple price changes. Next, excluding the cells already judgmentally selected, we took a comprehensive list of all other cells that had multiple 8-digits Census codes within them and randomly sampled 32. Of these, 14 were found to be suitable for the study. The detailed price information of the items of these cells was also obtained. The 23 total cells (7-digit) that are used for this project, along with the respective 8 digit levels, are shown in Appendix A.

3.2 Simulation Method

We conducted a simulation study of missing prices on the sampled cells. The simulation treated each cell independently. Item details for the month of June through August of 2004 were obtained.

All items with missing or estimated prices as well as items that had zero weight were removed from the data. Every item within data sets had a valid price from June to August 2004 and we considered these groups of items to be complete cells. For each cell, we calculated an index at the 8-digit cell level for August, using the following calculation methods:
Cell Index (August 2004) =

$$\frac{CellAggregate_{(August)}}{\sum w_{ij}} = \frac{\sum\left(w_i\ x\ \frac{P_{i,t}}{P_{i,b}}\right)}{\sum w_{ij}} = \left(\frac{\sum w_{ij,b} r_{i,t}}{\sum w_{ij,b}}\right)(100)$$

Where,

$CellAggregate_{(August)}$ = aggregates of all items in 8-digit cell level,

$\sum$ = the sum over all responding items in the cell,

$p_{i,t}$ = price of item *i* in August, 2004

$p_{i,b}$ = price of item *i* in July, 2004 (we use as the base period).

$r_{i,t} = p_{it} / p_{ib}$ = long term relative of item *i* in the month of August,

w_{ij} = the item i weight in cell j in July, 2004 (we use as the base period).

We considered these to be the ‘real’ index values. This real index at the 8-digit level was used to compare the estimated 7-digit and 8-digit indexes to test the effectiveness of the two relatives when estimating missing prices. Note that we use real index at 8-digit cell level because at this time we are only interested in knowing what the result of our cell relatives will be at this particular cell level.

Predetermined percents of items are randomly omitted from cells and defined as items with missing prices for the computation of the estimates of cell relatives at 7-digit and 8-digit level. The selected items are used to estimate the price relatives of the missing items. This process is repeated multiple times.

The following procedure is how the items with missing prices have been calculated by price relative method. In our study, prices of item i=4 in cell j has been removed (missing) for the month of August. This cell, say 8-digit NAICS cell j also contains three other items (i=1, 2, 3):

Estimated Price Relative of item 4, $i_{4j,(\text{August})}$ =

$$\frac{\sum_{i=1}^{n}(w_{ij}\times \mathrm{I}_{ij})}{\sum_{i=1}^{n} w_{ij}}$$

Estimated Cell Index for cell j =

$$\frac{\sum(w_{1j}\times \mathrm{I}_{1j}+w_{2j}\times \mathrm{I}_{2j}+w_{3j}\times \mathrm{I}_{3j}+w_{4j}\times i_{4j})}{\sum_{i=1}^{N} w_{ij}}$$

Where,

N = 4, (total number of items in cell j)

n = 3 (number of items with good price)

w_{ij} = the item i weight in cell j,

i_{4j} = Estimated Price Relative of item 4 in cell j

I_{ij} = the item i Price Relative in cell j in August.

The price relative of item i=4 derived from the above calculation is used as the estimated price change in the calculated month under the assumption that the previous month price relative for this item is 1 or 100 (in percent bases points).

We also calculated *percent good weight*, which is the sum of the weights of items with good prices in a cell compared to the total cell weights (sum of all the items weight in a cell):

Percent good Weight =

$$\frac{\text{sum of the weights of items with prices in cell j}}{\text{sum of the weights of all the items in cell j}}$$

$$= \frac{\sum_{i=1}^{\eta} w_{ij,b}}{\sum_{i=1}^{N} w_{ij,b}} \text{ x } 100.$$

Where,

η = Number of items in cell j with good prices
N = Total number of items in cell j
$w_{ij,b}$ = Weight of item i in cell j at time b.

Our study began by removing varying percentages of the items within a given cell. We started by removing 5% (Trial 1), then 15% (Trial 2), and continued incrementally by 10% until reaching 75% (Trial 8).

For each trial, the following procedure was simulated 100 times:

1. The specified percentage of items was removed randomly from the cell by Simple Random Sample SRS[1];
2. The items remaining were used to calculate an estimated 7 digit relative and an estimated 8 digit relative for each 8 digit code within the cell;
3. The items which were removed retained their July price but had their August price estimated using the 7- and 8-digit NAICS cell relatives from the previous step;
4. The items, which were randomly removed and had their August price estimated, are merged back into the cell;
5. Once the items, which had been removed, are merged back into the cell, two types of index estimates were calculated (one using the prices calculated by the estimated 7-digit NAICS cell relative method and the other using the prices calculated by the estimated product level relative method).
6. We repeated this process for each and every cell in our sample.

3.3 Computed Statistics:

The simulation produced two estimated indexes– one using the 7-digit cell relative method to estimate missing prices and the other using the product relative-8-digit cell for each of the 100 simulations within each trial. Note that we are using the actual sample as our population here. This means that we resample k times from our original sample. The variable chosen for analysis was the absolute difference or also known as absolute deviation between the two estimated indexes and the actual August index (index which includes all items and their actual prices) at the 8-digit level.

$$AD^{8} = abs\left(Index\,\Pr oductLevel_{8digit-cell} - \mathrm{Re}\,al\,\Pr oductLevelIndex\right)$$

$$\Delta_8$$

$$= \left| \left(\overbrace{\frac{\sum w_{hij,b} r_{i,t}}{\sum w_{hij,b}}}^{Estimated} \right)(100) - \left(\overbrace{\frac{\sum w_{hi,b} r_{i,t}}{\sum w_{hi,b}}}^{Complete-data} \right)(100) \right|$$

$$AD^{7} = abs\left(Index\,\Pr oductLevel_{7digit-relative} - \mathrm{Re}\,al\,\Pr oductLevelIndex\right)$$

$$\Delta\ 7 = \left| \left(\overbrace{\frac{\sum w_{hij,b} r_{i,t}}{\sum w_{hij,b}}}^{Estimated} \right)(100) - \left(\overbrace{\frac{\sum w_{hi,b} r_{i,t}}{\sum w_{hi,b}}}^{Complete-data} \right)(100) \right|$$

Absolute Difference (AD) is our chosen measure of error between the actual calculated index, θ_{gh} from our original sample with all the items' prices, and the estimated index, $\hat{\theta}_{gh,\,j}$ calculated from the same sample but this time some of the items in the cell (stratum), h have their items removed as missing.

Thus the above formula can be expressed as:

$$\Delta_{g,k} = \left| \ \hat{\theta}_{gh,\,j,k} - \theta_{8h} \ \right|$$

Where,

$$\theta_{8h} = \left(\overbrace{\frac{\sum w_{hib} r_{i,t}}{\sum w_{hi,b}}}^{Complete-data} \right)(100),$$

[1] The number of items within the cell is multiplied by the specified percentage and the results were rounded up the next whole number.

7,8 Absolute difference between seven, eight digits cell levels and the real product code (8-digit cell level) relatives, respectively

$$\hat{\theta}_{ghk} = \overbrace{\left(\frac{\sum w_{hij,b} r_{i,t}}{\sum w_{hij,b}}\right)}^{Estimated}(100),$$

$$= \hat{\theta}_{ghk} = \overbrace{\left(\frac{\sum w_{hkj} I_{hkj}}{\sum w_{hkj}}\right)}^{\text{Imputed weighted relatives}}(100)$$

$w_{hi,b}$ = Item i weight from previous month b in cell h

w_{hkj} = Cell h weight recalculated after every j trial

g = The index level, that is 7-digit or 8-digit cell,

h = Cell (Stratum), (h= 1,2…36),

j = Trial, (j=1, 2…10),

k = Independently repeated sample selection k number of times, (k=1,2, …100).

And,

$$I_{hkj} = \left(\frac{\sum_{i=1}^{n-c} w_{hkji} r_{hkji} + \sum_{i=1}^{c} w_{hkji} \hat{r}_{hkji}}{\sum_{i=1}^{n} w_{hkji}}\right), \text{ c} \geq 1 \text{ (missing items)}$$

$$\hat{r}_{hkj(Estimate)} = \left(\frac{\sum_{i=1}^{n-c} w_{hkji} r_{hkji}}{\sum_{i=1}^{n-c} w_{hkji}}\right), \text{ n - c} \geq 1 \text{ (items with current prices)}$$

I_{hkj} = Imputed cell h relative of K replicate at trial j

$\hat{r}_{hkj(Estimate)}$ = Estimated item price relative

We calculated the Mean Absolute Difference (MAD) of $\Delta_{g,k}$ after simulating it for k times:

$$\overline{\Delta}_{g,h} = \frac{\sum_{k=1}^{100} \left| \hat{\theta}_{gh,j,k} - \theta_{gh} \right|}{K_j}, \qquad K_j = 100$$

(number of times simulated)

So, for the MAD at 7-digit index level of cell h we have:

$$\overline{\Delta}_{7,h} = \frac{\sum_{k=1}^{100} \left| \hat{\theta}_{7h,j,k} - \theta_{8h,} \right|}{100},$$

And for the MAD at 8-digit index level of cell h we computed it as:

$$\overline{\Delta}_{8,h} = \frac{\sum_{k=1}^{100} \left| \hat{\theta}_{8h,j,k} - \theta_{8h,} \right|}{100},$$

We examined the mean absolute difference with respect to two variables: the number of items remaining at the 8-digit level and the percent good weight remaining at the 8-digit level. With the mean differences examined by the number of items remaining as well as the percent good weight, we were able to look for trends at the individual NAICS cell level and overall (all the 36 NAICS) cells combined.

A small value of MAD indicates small relative error between our estimated index and the real index of the sample. Hence the use of index level with a smaller value of MAD would be a better choice in estimating item relatives for a typical cell that has that level of item availability.

4. RESULTS:

a. Overall Cells

Overall results were obtained by combining the data for all 23 cells.

The graph below (figure 1) shows mean absolute difference as the number of items on the product level increases. According to the graph, the 7-digit level relative proves to be a better estimator of missing prices when few items are at the 8-digit level. On average, once an 8 digit level cell has approximately 12 or more items the product relative becomes the better estimator for the missing price. The cause of this, we believe, is the homogeneity of the 8 digits cells as the items increase within the cells. Neither is clearly better when they are between 7 and 13 items, results are neutral.

Figure 1.

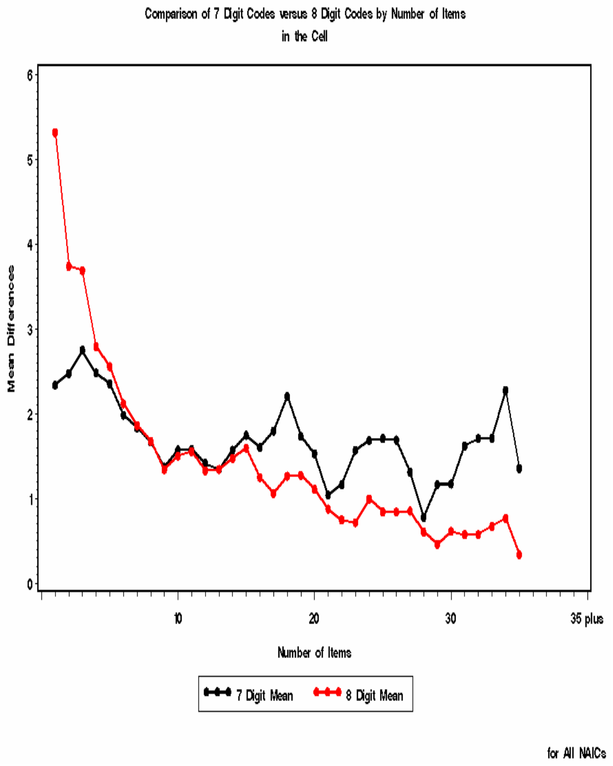
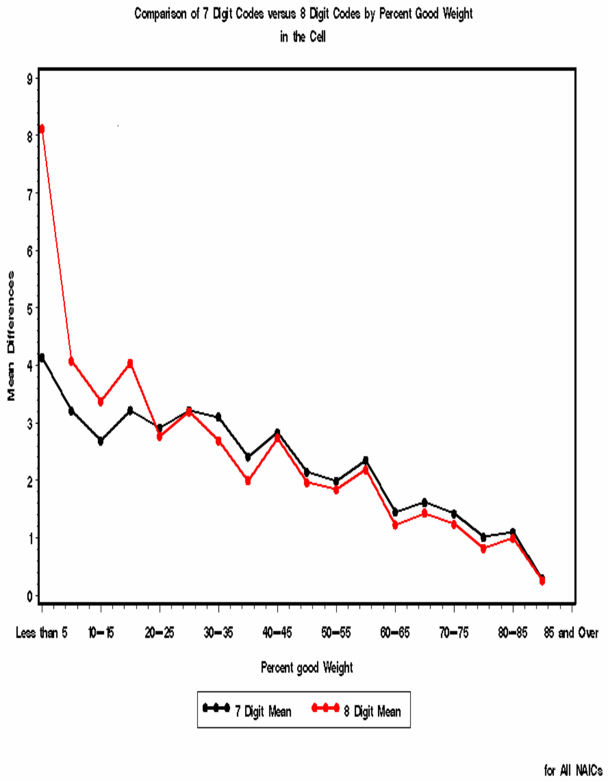


Figure 2.



Next we examined the effect the percent good weight, which is the weight associated with items with current prices in the cell. This percent good weight remaining at the product level is examined to determine the effect it had on the two estimation methods, (see figure 2). This is a graph of the mean absolute difference by categories of percent good weight remaining at the 8-digit level. Percent good weight appears to have a similar effect on the estimates by the two estimation methods as the number of items had. When only a small percentage of good weight remains at the product level, using the 7-digit relative to estimate the missing prices produces an index that is closer to the true index. When percent of good weight is at 25% or higher, the product relative produces better estimates closer to the true value.

b. Individual Cells

We then looked at each of the 23 cells individually to see if there were notable trends among certain groups of cells.

Looking at the individual NAICS cells, the picture is a little bit murky in contrast with the clear picture painted by the overall result where we have all the sampled NAICS cells combined or aggregated. But this might be due to an aberration with a few NAICS cells, whose price changes exhibited haphazard pattern (especially industries at the service sector of the economy).

The effect of these individual cells on the aggregate is muted or subdued partly, because most of the NAICS cells exhibiting this chaotic behavior have very little price change. In effect they have a very narrow range between the maximum and the minimum mean difference between our estimated and the real values of cell relatives. Often, this range is between 0.00 and 0.9 (in percentage points).
The individual sampled NAICS cells are categorized into groups of less than 12 and more than 12 items per cell. The table1 below shows the breakdown of the 23

NAICS by method of imputing the cell relatives and by number of items.

Table1. Number of NAICS cells per method

Number	Cell category	Less than 12 items	Greater than 12 items
1	7-digit cell relative better	11	2
2	8-digit cell relative better	2	4
3	Neutral (either of the methods)	2	10

Five NAICS cells have abnormal price pattern, which judging from their industrial origins, prices of items in these industrial groups tend to be cyclical.

In the above table we have seen that, although there are instances where there is a deviation from norm, the result shows an intriguing instance where the 7-digit cell relative fare better when the number of items in the cell is fewer than 12. However, it is less obvious which result is better when the number of items exceeds 12. While the 8-digit cell relative method was a better estimate in twice as many cells when there was a difference, the few number of cells with a measurable difference makes the result somewhat unclear.

6. CONCLUSION

Even though results varied from cell to cell, the overall results suggest that a cut off for using the 8-digit Census code cell relative for estimating missing prices is when there are more than 12 items with nonzero price change as well as at least 25 percent good weight available at the product level. If these two conditions are not met, it is advisable to use the 7-digit Census code cell relative to estimate the prices. For optimal price estimates, it might be advisable to look at each cell specific individually, although it would be quite time intensive.

It is very important to stress that further work needs to be done in order to reach a definitive conclusion. We plan to expand this study to examine more cells, including looking at various time periods.

And calculating the variances of the imputed price relatives (point estimates) of the missing prices.

ACKNOWLEDGMENTS

The author would like to thank Jennifer S. Willen and Mandy Webb for providing valuable information in researching particular instances of NAICS cells structures.

Special thanks to Richard J. Pratt for his support on this project.

7. REFERENCES

Brick, J.M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Bureau of Labor Statistics (2003), Handbook of Methods, *Washington. DC, http://www.bls.gov/opub/hom/home.htm*

Cochran, W.G. (1977), Sampling Techniques, Third Edition, John Wiley & Sons.

Kalton, G. (1983), Compensating for Missing Survey Data, Ann Arbor, MI: *Institute of Social Research, University of Michigan.*

Kish, L. (1965), Survey Sampling, New York: John Wiley & Sons, Inc.

Kish, L. (1987), Statistical Design for Research, New York: John Wiley & Sons, Inc.

Producer Price Index Concepts and Methods (2005), Bureau of Labor Statistics, Washington, DC.

SAS/STAT User’s Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

Williams, R.L. and Chromy, J.R. (1980), “SAS Sample Selection Macros,” Proceedings of the Fifth Annual SAS Users Group International Conference, 5, 392-396.