Variance Estimations for International Price Program Indexes

Te-Ching Chen[1], Patrick A. Bobbitt[2], James A. Himelein Jr.[2], Steven P. Paben[2],
MoonJung Cho[2], Lawrence R. Ernst[2]
U.S. Bureau of Labor Statistics/Digital Management, Inc.[1]
U.S. Bureau of Labor Statistics[2]

Abstract

The International Price Program (IPP) collects data on United States' trade with foreign nations and publishes monthly indexes on import and export prices of U.S. merchandise and services. Recently, the IPP evaluated different variance estimation methods such as Taylor Series Linearization, bootstrap, jackknife, and BRR, for their applicability to the IPP. We constructed an artificial universe of monthly price changes for items constructed from 13 years of IPP historical data. We then compared the bias and stability of the variance estimation methods for month-to-month, annual, and long-term price changes by drawing from the universe 1000 samples in various merchandise strata.

KEY WORDS: Variance estimation; Bootstrap; Jackknife; Balance Repeated Replication (BRR); Certainty Sampling Units; Taylor Series.

1. Introduction

The International Price Program (IPP) of the Bureau of Labor Statistics (BLS) collects data on United States' trade with foreign nations and publishes monthly indexes on the changes in import and export prices for both merchandise and services. The purpose of the IPP Variance Project is to study the estimated variance of the index using different variance estimation methods. The variance estimation methods we studied in this project are a simplified Taylor Series linearization method that is currently being used, an alternative Taylor Series linearization formula, and resampling methods of Jackknife, Bootstrap and Balance Repeated Replication (BRR). The study consisted of creating an artificial universe of price relatives from which we drew 1000 simulated samples. Then, the variance of the 1000 simulated samples was calculated and treated as the benchmark to which we compare each variance estimation method.

2. IPP Sampling and Index Estimation

2.1 Background

The IPP, as the primary source of data on price change in the foreign trade sector of the U.S. economy, publishes index estimates of price change for internationally traded goods using three different classification systems - Harmonized System (HS), Bureau of Economic Analysis End Use (BEA) and North American industry classification system(NAICS). In this study, we only study the variances of the import price indexes.

The major price programs at the BLS use the following general approach for calculating price indexes. A sample market basket of items is drawn to be representative of the universe of prices being measured. Prices for the items in that market basket are then collected from month to month. Using an index methodology that holds quantities fixed, price indexes are derived measuring pure price change as distinct from changes in the product mix.

The target universe of the import price index consists of all goods and services purchased from abroad by U.S. residents. Ideally, the total breadth of U.S. trade in goods and services in the private sector would be represented in the universe. Items for which it is difficult to obtain consistent time span for comparable products, however, such as works of art, are excluded. Products that may be purchased on the open market for military use are included, but goods exclusively for military use are excluded.

2.2 Sampling in the International Price Program

The import merchandise sampling frame used by the IPP is obtained from the U.S. Customs and Border Protection. Because shippers are required to document nearly all trade into and out of the U.S., IPP is able to sample from a fairly large and detailed frame. The frames contain information about all import transactions that were filed with the U.S. Customs and Border Protection during the reference year. The frame information available for each transaction includes a company identifier, usually the Employer Identification Number, Harmonized Tariff number, the detailed product category of the goods that are being shipped and the corresponding dollar value of the shipped goods.

IPP divides the import merchandise universe into two halves referred to as panels. A sample for one panel is selected each year and sent to the field offices for collection, so the universe is fully resampled every two years. The sampled goods are priced for approximately five years until they are replaced by a fresh sample from the same panel. As a result, each published index is based upon the price changes of items from up to three different samples.

Each panel is sampled using a three stage sample design. The first stage selects establishments independently

proportional to size (pps) within each broad product category (stratum) identified using HS classifications. The measure of size is the total trade dollar value of the establishment within the stratum.

The second stage selects detailed product categories (i.e., classification groups) within each establishment-stratum using a systematic pps design. The measure of size is the relative dollar value adjusted to ensure adequate coverage for all published strata across all classification systems (HS, BEA and NAICS), and known non-response factors such as total company burden and frequency of trade within each classification group. Each establishment-classification group (or sampling classification group, SCG) may be selected multiple times. The number of items a SCG is selected is called the number of quotes requested.

In the third and final stage, the BLS Field Economist, with the cooperation of the company respondent, performs the selection of the actual item for use in the IPP indexes. Ideally, the respondent will be able to identify a list of items along with each items total trade dollar value within each selected sampling classification group. The field economist in conjunction with the respondent will then complete further stages of sampling until one item for each quote is selected. This process is called disaggregation. This process is done with replacement, so the same item may be selected more than once.

2.3 Index Estimation

IPP uses the items that are initiated and re-priced every month to compute its indexes of price change. These indexes are calculated using a modified Laspeyres index formula. The modification differs from the conventional Laspeyres in that the IPP 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. However, 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, and a suitable means for handling items that are not traded every calculation month. The modified fixed quantity Laspeyres formula used in the IPP is as follows:

$$\begin{aligned} LTR^t &= \left(\frac{\sum w_i^0 r_i^t}{\sum w_i^0}\right)(100) \\ &= \left(\frac{\sum w_i^0 r_i^t}{\sum w_i^0 r_i^{t-1}}\right)\left(\frac{\sum w_i^0 r_i^{t-1}}{\sum w_i^0}\right)(100) \\ &= (STR^t)(LTR^{t-1}) \end{aligned}$$

Where:

$$\begin{aligned} LTR^t &= \text{long-term relative of a collection of items at time } i \\ p_i^t &= \text{price of item } i \text{ at time } t, \\ q_{i,0} &= \text{quantity of item } i \text{ in base period 0}, \\ w_{i,0} &= (p_{i,0})(q_{i,0}) \\ &= \text{total revenue of item } i, \text{ in base period 0}, \\ r_i^t &= p_i^t/p_{i,0} \\ &= \text{long-term relative of item } i \text{ at time } t, \\ STR^t &= \left(\sum w_{i,0} r_i^t\right) / \left(\sum w_{i,0} r_i^{t-1}\right) \\ &= \text{the short-term relative of a collection of items } i \text{ at time } t. \end{aligned}$$

For each classification system, IPP calculates its estimates of price change using an index aggregation structure (i.e. aggregation tree) with the following form:
Upper Level Strata
Lower Level Strata
Classification Groups (CG)
Weight Group (i.e., Company-index Classification Group)
Items

The classification groups' level is the highest common level for all of the classification systems. Each classification system has a different set of lower level and upper level strata. The SCGs' weights equal to the sum of the item weights. However, IPP uses fixed aggregation weights for Classification Groups and upper levels.

3. Variance Estimation Methods

In multi-stage samples, one typically only estimates the variance given the first stage of sampling (Wolter(1985)). This implies that units selected with certainty should have their variability accounted for at a subsequent stage of sampling. In our sample design, it is possible to have both certainty and probability (or non-certainty) units selected at each stage.

Index or percent change estimates involve ratios which are not linear functions. One method of estimating variances of a non-linear function is to linearize the interested non-linear estimators using a Taylor series approximation. Another ways to estimate the variance of a function are resampling methods. The basic idea of resampling methods is to calculate the estimate of interest from the full sample as well as a number of subsamples or replicates. The variation among the subsample or replicate estimates is used to estimate the variance of the full sample. We studied three commonly used resampling methods: Jackknife, Bootstrap, and Balanced Repeated Replication (BRR). For each of these methods, we tried at least two variations.

3.1 Taylor Series Linearization Method

The current variance estimates are only calculated for the sample variance of 12-month changes using a Taylor Series linearization method (TAYLOR). The TAYLOR

treats establishments selected with certainty in the first stage the same as establishments selected with probability. Since it is customary in multi-stage samples to calculate variance estimates at the first stage of probability selection, we also calculated variance estimates using an alternative Taylor series linearization method (TAYLORC) that should account for the variability of second stage units for certainty establishments. However, the results for TAYLORC were not any better than TAYLOR. Therefore, we will only discuss the results for the TAYLOR method in this paper.

The TAYLOR is currently used to estimate the variance of lower level strata. This method pools all items within a lower-level stratum to obtain a variance estimate. For this study, we extended the pooled estimator to the upper level strata. We also extended it to estimate the variance for the STRs.

Recall that we assume the sample design is simplified to two stages and all of the establishments are non-certainties. The first stage selects n_h establishments without replacement. For each of the sampled establishments represented by $\{j\}$, a sample of m_{hj} items is drawn with replacement. The estimator of interest is the ratio that is defined:

$$\hat{R}_h = \frac{\hat{Y}_h}{\hat{X}_h} = \frac{\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}} w_{hji} r_{hji}^{(t)}}{\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}} w_{hji} r_{hji}^{(t-12)}}. \quad (1)$$

where $\left\{r_{hji}^{(t)}, r_{hji}^{(t-12)}\right\}$ is the set of long-term relatives at time t and $t-12$ and
w_{hji} is item weights for item i of establishment j in stratum h.

The Taylor series estimator of the variance of $\hat{R}_h$ is

$$Var(\hat{R}_h) = \hat{R}_h^2 \left[\frac{Var(\hat{Y}_h)}{\hat{Y}_h^2} + \frac{Var(\hat{X}_h)}{\hat{X}_h^2} - 2\frac{Cov(\hat{Y}_h, \hat{X}_h)}{\hat{X}_h \hat{Y}_h}\right] \quad (2)$$

Let

$$w_{hj} = \sum_{i=1}^{m_j} w_{hji}, \text{ and}$$

$$\hat{z}_{hj} = R_h \sum_{i=1}^{m_j} \frac{w_{hji}}{w_{hj}} \left(\frac{r_{hji}^{(t)}}{\sum_{j'=1}^{n_h}\sum_{i'=1}^{m_{hj'}} w_{hj'i'} r_{hj'i'}^{(t)}} - \frac{r_{hji}^{(t-12)}}{\sum_{j'=1}^{n_h}\sum_{i'=1}^{m_{hj'}} w_{hj'i'}^{(t-12)}} \right)$$

Then

$$Var(\hat{R}_h) \doteq \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} \left(w_{hj}\hat{z}_{hj}\right)^2. \quad (3)$$

This approach is detailed in Himelein (2006).

3.2 Jackknife Method

The jackknife method generally consists of splitting the total sample into G disjoint and exhaustive Primary Sampling Units (PSUs), then dropping out a specified number of PSUs in turn, and estimating the parameter of interest from the remaining units each time. The variability among the replicate estimates is then used to estimate the variance of the full-sample estimator.

In our study, we used a "delete-one" stratified jackknife method in which exactly one PSU was dropped within a stratum for each replicate. The general form of the variance estimator we used is:

$$Var(\hat{R}_h) = \frac{n_h - 1}{n_h} \sum_{k=1}^{n_h} \left(\hat{R}_{hk} - \hat{R}_h\right)^2,$$

Where

- $\hat{R}_{hk}$ is the estimate of $\hat{R}$ with the k^{th} unit dropped out in stratum h,
- n_h is the number of units and replicates dropped out in stratum h, which implies there would be $N = \sum_{h=1}^{H} n_h$ replicates for the estimate of interest.

The delete-one estimates denoted $\left\{\hat{R}_{hk}\right\}$ are derived using a set of adjusted weights. Let $k' = 1, \ldots, g_h$ represent the random groups of each variance stratum h and $\{w_k\}$ be the sum of the index weights within each random group of the variance stratum. The weight adjustment for the units of the variance stratum is then the following:

(1) $w^*_{(hk),hk'} = 0$, for $k' = k$
(2) $w^*_{(hk),hk'} = w_{hk}\left(\frac{n_h}{n_h - 1}\right)$ for $k' \neq k$, where n_h is the total number of PSUs in the variance stratum.

The variance strata and variance PSUs are assigned based on the sample design of the survey. Again, we assume the sample design is simplified to two stages. The variance PSUs should be assigned at the first stage of probability selection. For non-certainty establishments, the variance strata are equal to the sampling strata and each establishment is a variance PSU. If a non-certainty establishment has items in more than one sampling classification group(SCG) in a particular sampling stratum, all of the items belonging to that non-certainty establishment are deleted at the same time and the item weights for remaining items of this particular sampling stratum are adjusted separately by the PSU counts of each classification group. For certainty establishments, the variation for non-certainty SCGs is from the second stage selection of the SCGs but for certainty SCGs the variation is from the selection of the items. However, there is only one item available to be selected in each non-certainty SCG and at least one item for certainty SCGs. Therefore, the majority of the variation of certainty establishments is the item selection. Hence, for certainty establishments,

each establishment is a variance stratum and the items of the establishment are the variance PSUs.

For those SCGs of certainty establishments with only one item selected, a collapsed stratum process is applied to "pair up" variance strata. If there are an odd number of cases greater than one for the CG, one of the "collapsed" variance strata will have three items. We also did the weight adjustments based upon the deleted PSU weights, but both theoretical and simulation results show they are close to each other.

3.3 Bootstrap Method

For stratified designs, a direct extension of the standard bootstrap is to apply it independently for each stratum. This methodology is often referred to as the naive bootstrap. Since the naive bootstrap variance estimator is inconsistent in the case of bounded sample sizes, several modified bootstraps have been proposed. One of those methods is the rescaling bootstrap proposed by Rao, Wu, and Yue (1992). The bootstrap method they proposed was designed to handle stratified multi-stage samples. This method takes a SRSWR (simple random sampling with replacement) of PSUs and applies a scale adjustment to the final survey weights to account for the variability of units selected at subsequent stages.

Different bootstrap methods were studied in this project. Here, we only present the method we think would best capture all of the variability of the IPP sample design. For more information on the other bootstrap methods, please refer to Bobbitt et al.(2007).

For the bootstrap rescaling method, we created a separate variance strata for certainty establishments similar to the jackknife method. However, here we created an additional set of variance strata to distinguish between certainty and non-certainty SCGs selected within certainty establishments.

Let S_h be the set of all sampled items from sampling stratum h which is partitioned into three groups as follows:

$$h_1 = \left\{ \begin{array}{l} \text{Set of items in } S_h \text{ selected from} \\ \text{non-certainty establishments} \end{array} \right\}$$

$$h_2 = \left\{ \begin{array}{l} \text{Set of items in } S_h \text{ selected from} \\ \text{non-certainty SCGs within certainty} \\ \text{establishments} \end{array} \right\}$$

$$h_3 = \left\{ \begin{array}{l} \text{Set of items in } S_h \text{ selected from certainty} \\ \text{SCGs within certainty establishments} \end{array} \right\}$$

We attempted to "pair up" single unit variance strata within sampling stratum h. Each of these new variance strata contained at least two units. In the case of an odd number of units greater than one, the new variance stratum will have three units.

In this method, we select:

$$n^b_{h_p} = \left\{ \begin{array}{ll} n_{h_p} - 1 & n_{h_p} > 1 \\ 1 & n_{h_p} = 1 \end{array} \right\}, \text{ where } \begin{array}{l} p = 1, 2, \text{ or } 3 \\ \text{partition of } S_h, \end{array}$$

units in each bootstrap sample. If we substitute w_{h_pji} for w_{hji} and $n^b_{h_p}$ for n^b_h, then the bootstrap weight is:

$$w^b_{h_pji} = \left\{ \begin{array}{ll} w_{h_pji}\left(\frac{n_{h_p}}{n_{h_p}-1}\right) m^b_{h_pji} & \text{for } n_{h_p} > 1 \\ w_{h_pji} & \text{for } n_{h_p} = 1 \end{array} \right\}$$

Drawing 150 bootstrap samples of size n^b_h, the variance estimator is:

$$V\left(\hat{R}_h\right) = \frac{1}{150}\sum_{b=1}^{150}\left(\hat{R}^b_h - \hat{R}_h\right)^2, \tag{4}$$

where $\hat{R}_h$ is the price relative of interest for stratum h using the original sample.

3.4 Standard BRR (BRR)

The standard BRR design assumes that a population of PSUs are able to be grouped into H strata with two PSUs selected per stratum using with replacement sampling. Then, K replicate half-sample estimates are formed by selecting one of the two variance PSUs from each stratum based on a Hadamard matrix and then using only the selected variance PSU to estimate the parameter of interest. The sampled weights for the selected units are doubled to create a set of replicate weights from which to calculate replicate estimates. In order to obtain a fully balanced design, the number of replicates used needs to be a multiple of four greater than the number of strata.

The formula for calculating the variance of the relative of interest in stratum h is then simply:

$$V\left(\hat{R}_h\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{R}_{hk} - \hat{R}_h\right)^2 \tag{5}$$

where

$\hat{R}_h$ is the estimated price relative for stratum h using the original full sample,
$\hat{R}_{hk}$ is the estimated price relative for stratum h based on the k^{th} replicate,
K is the total number of half-sample replicates,

As in the other variance procedures, the variance strata should be assigned based on the sample design of the survey. Since BRR requires two PSUs per stratum and the IPP sample design has more than two PSUs per stratum, we artificially created the variance strata within each sampling stratum. Each certainty establishment is a separate stratum. While variance strata for non-certainty establishments are assigned two establishments per stratum. In order to prevent using many different size Hadamard matrices, we used only a handful of them based on groupings for different sizes of strata.

Within each variance stratum, two variance PSUs are created. The variance PSUs should also be assigned to be consistent with the sampling design. For certainty establishments, items are sorted by descending probability of selection and alternately assigned to one of two

variance PSUs. Certainty establishments with even number of items selected for an SCG should have the quotes evenly split between the two variance PSUs, while certainty establishments with an odd number of quotes selected for an SCG should have one variance PSU assigned only once more than the other variance PSU. For non-certainty establishments, each establishment is alternately assigned to one of two variance PSUs and all items within each non-certainty establishment are assigned to the same variance PSU.

4. Simulation Model

One method of assessing different sample variance estimation methodologies is to compare each method to the "true" measure of the sample variance. For our universe of the item relatives, we pooled the company-classification groups of the 1000 samples drawn from the 30^{th} Import Sample (July 2002 - June 2003). Due to limitations of the sampling frame, we were forced to assume that the same item would be selected during disaggregation for each company-classification group. For company-classification groups that are selected more than once, multiple item relatives were created to equal the maximum number of selections for the company-classification group in the 1000 samples. The repeated simulations allow for the calculation and comparison of a number of evaluative statistics such as bias, stability, and confidence interval coverage.

You may recall that if $X \sim F(x)$, then two things are true.

$$F(x) \sim Unif(0,1) \quad (6)$$

$$U \sim Unif(0,1) \Rightarrow F^{-1}(u) \sim X \quad (7)$$

This suggests we can get an estimate of the empirical distribution of item STRs for a given CG and month by estimating the percentiles of the historical price data for a given CG and month. Then, we can obtain an estimate of the $F(x)$ for each CG and month using linear interpolation between data points. According to the historical data, 99.94% of more than 4 million not-imputed item STRs are between 0.5 and 2.

The sampling procedures for the simulation study follows the IPP sampling procedures described in 2.2 as closely as possible. For more detail information about creating the universe and the empirical distribution, please see Cho et al.(2007).

5. Simulation Results

5.1 Overview

We estimated the variances with different variance estimation methods. For each method, we estimated the variance of the STRs and LTRs from Month 1 to Month 36 and annual changes from Month 12 to Month 36. We are primarily interested in the estimated variances for the 31 published two-digit HS strata for STRs and annual changes. In the presented analysis, we focus on the average variance estimates for the annual changes up to Month 24 that were calculated for each of the different methods and strata.

5.2 Analysis Formulas

We compare the simulation results in terms of relative bias, stability, and coverage rate of the 95% confidence interval for the average of the annual changes over months for each method.

For each two-digit HS stratum h, let us define y_i be the full vector of entire sample i where $i = 1, \ldots, 1000$, $\hat{\theta}_{hi} = \hat{\theta}_h(y_i)$ and define

$$\bar{\hat{\theta}}_{h.} = \frac{1}{1000} \sum_{i}^{1000} \hat{\theta}_{hi} \quad (8)$$

$$\tilde{V}_{h.} = \frac{1}{1000-1} \sum_{i}^{1000} \left(\hat{\theta}_{hi} - \bar{\hat{\theta}}_{h.} \right)^2 \quad \text{and} \quad \tilde{\sigma}_{h.} = \sqrt{\tilde{V}_{h.}}$$

As we do not have a true variance, for each of the two-digit HS stratum h, we use $\tilde{V}_{h.}$ and $\tilde{\sigma}_{h.}$ as our population variance and standard deviation. We compare the standard error estimations with $\tilde{\sigma}_{h.}$.

Let $\hat{\sigma}_{mhi}$ the standard error estimator of a two-digit HS stratum h of sample i for the variance estimation method m. The relative bias of an interested variance estimation method is calculated as

$$\text{Relative Bias} = \frac{\left(\frac{1}{1000} \sum_{i}^{1000} \hat{\sigma}_{mhi} \right) - \tilde{\sigma}_{h.}}{\tilde{\sigma}_{h.}} \quad (9)$$

and the stability is

$$\tilde{\sigma}(\hat{V}_{mhi}) = \sqrt{\frac{1}{1000-1} \sum_{i}^{1000} \left(\hat{V}_{mhi} - \bar{\hat{V}}_{mh.} \right)^2} \times 100. \quad (10)$$

where $\bar{\hat{V}}_{mh.}$ is the average of the 1000 variance estimations for the method. We use the percentage of the biases and the stability as are interested in percent change estimates.

The coverage rate are calculated as

$$\hat{c} = \frac{1}{1000} \sum_{i}^{1000} I \left\{ \bar{\hat{\theta}}_{h.} \in \left(\hat{\theta}_{mhiL}, \hat{\theta}_{mhiU} \right) \right\}. \quad (11)$$

Where $I = 1$ if $\bar{\hat{\theta}}_{h.} \in \left(\hat{\theta}_{mhiL}, \hat{\theta}_{mhiU} \right)$ or 0 otherwise. The $\hat{\theta}_{mhiU}$ and $\hat{\theta}_{mhiL}$ are the upper and lower bounds of confidence intervals of $\hat{\theta}_{hi}$. The $\bar{\hat{\theta}}_{h.}$ is defined in equation (8), the average of 1000 index estimates is used as the population or "true" index.

5.3 Average Standard Error Results for Annual Change

Table 1 lists the percentage of the average relative biases of the standard error, stability and 95% coverage rate for

Simulation Results for Average of 13 Annuals												
	BRR			Bootstrap			Jackknife			TAYLOR		
Stratum	RBias[a]	Stab[b]	CI[c]	RBias	Stab	CI	RBias	Stab	CI	RBias	Stab	CI
P02	0.11	0.132	96.3%	0.04	0.091	95.6%	0.27	0.175	96.6%	0.07	0.199	94.7%
P03	0.30	0.293	97.4%	0.06	0.175	95.3%	0.12	0.267	95.0%	0.33	1.896	96.5%
P07	0.00	41.622	89.3%	-0.18	31.304	84.4%	-0.11	50.683	82.6%	-0.15	45.003	79.8%
P08	-0.10	19.876	80.6%	-0.24	14.477	76.9%	-0.13	28.690	74.0%	-0.09	40.720	78.8%
P09	0.20	0.427	96.3%	0.08	0.358	95.4%	0.24	0.519	96.9%	0.15	0.476	96.3%
P20	0.18	0.135	98.2%	-0.03	0.088	94.9%	0.10	0.148	95.3%	0.46	1.050	96.5%
P22	0.19	0.006	96.9%	0.04	0.004	95.1%	0.18	0.008	96.2%	0.22	0.012	96.7%
P42	0.11	0.004	96.7%	-0.04	0.003	94.9%	0.18	0.006	97.4%	0.17	0.020	95.5%
P47	0.00	0.063	96.3%	-0.05	0.049	95.9%	0.08	0.106	97.4%	0.14	0.074	98.1%
P48	0.19	0.007	96.7%	0.10	0.004	96.6%	0.32	0.008	98.2%	0.29	0.017	98.0%
P49	0.08	0.008	96.1%	-0.09	0.005	93.6%	0.07	0.012	93.8%	0.17	0.026	92.7%
P61	0.12	0.001	97.6%	0.05	0.001	96.7%	0.18	0.001	98.2%	0.33	0.005	98.2%
P62	0.07	0.002	96.5%	-0.03	0.002	94.4%	0.11	0.003	97.1%	0.31	0.014	97.2%
P63	0.11	0.009	97.3%	-0.08	0.006	93.5%	0.04	0.012	95.3%	0.34	0.035	97.3%
P64	0.12	0.001	97.8%	0.10	0.001	97.3%	0.26	0.002	98.2%	0.28	0.007	97.1%
P68	-0.03	0.013	95.9%	-0.19	0.009	92.2%	-0.01	0.019	92.8%	0.02	0.019	92.0%
P69	0.03	0.018	97.4%	-0.14	0.012	93.6%	0.02	0.030	92.6%	0.02	0.073	92.9%
P70	0.15	0.013	96.9%	-0.10	0.007	92.5%	0.02	0.011	92.9%	0.16	0.031	95.5%
P72	0.30	0.064	98.1%	0.19	0.062	97.0%	0.09	0.059	94.2%	0.81	0.407	99.0%
P73	0.10	0.045	96.8%	-0.12	0.021	91.6%	-0.07	0.021	90.8%	0.42	0.799	96.9%
P74	-0.02	0.317	94.8%	-0.20	0.235	89.5%	-0.30	0.309	75.6%	-0.40	0.128	76.4%
P76	0.27	0.058	98.1%	0.06	0.040	95.6%	0.31	0.084	97.3%	0.17	0.104	96.5%
P82	0.01	0.017	96.6%	-0.19	0.010	91.5%	-0.09	0.017	90.6%	0.33	0.177	95.9%
P83	0.09	0.016	96.5%	-0.11	0.011	92.4%	0.08	0.025	93.0%	-0.07	0.023	90.8%
P87	0.21	0.008	97.2%	0.07	0.007	95.3%	0.15	0.009	95.4%	-0.23	0.007	79.3%
P88	0.09	0.002	98.1%	0.11	0.002	98.4%	0.27	0.004	99.2%	0.33	0.004	99.5%
P90	0.13	0.004	98.0%	0.01	0.003	97.6%	0.09	0.002	97.5%	0.18	0.013	97.2%
P91	0.19	0.017	97.7%	0.04	0.011	96.5%	0.18	0.019	97.4%	0.37	0.063	98.2%
P94	0.12	0.002	96.6%	0.03	0.001	95.1%	0.17	0.003	97.1%	0.19	0.005	96.1%
P95	0.30	0.006	94.7%	0.13	0.004	93.5%	0.42	0.008	96.1%	0.38	0.021	95.3%
P96	0.15	0.033	96.7%	-0.05	0.026	95.0%	-0.03	0.036	93.1%	0.10	0.050	95.1%
Ave 1[d]	0.12	2.039	96.1%	-0.02	1.517	93.8%	0.10	2.622	93.8%	0.19	2.951	93.9%
STD 1[e]	0.10	8.163	3.3%	0.11	6.106	4.1%	0.15	10.294	6.0%	0.23	10.676	6.3%
Ave 2[f]	0.13	0.059	96.9%	-0.01	0.043	94.7%	0.12	0.066	94.9%	0.21	0.198	94.9%
STD 2[g]	0.09	0.107	0.9%	0.10	0.082	2.1%	0.15	0.118	4.4%	0.22	0.410	5.1%

[a]Relative Bias
[b]Percentage of the Stability
[c]Coverage Percentage of the 95% confidence interval
[d]Average over the strata
[e]Standard Error over the strata
[f]Average over the strata, without P07, P08
[g]Standard Error over the strata, without P07, P08

Table 1: Simulation Results for Average of 13 Annual Changes

the 31 interested two-digit HS strata for each method. All of the averages are based on simulation results of 13 annual changes. Comparison results of relative bias and stability will be described in Section 5.3.1 and Section 5.3.2 for confidence interval comparisons.

5.3.1 Bias and Stability

Figure 1 shows the simulation results for the average relative bias for each two-digit HS stratum. All of the methods underestimated P07 (Edible vegetables, roots, and tubers), P08(Edible fruit and nuts; peel of citrus fruit or melons) and P74 (Copper and articles thereof). Strata P07 and P08 are highly seasonality items and have shown wide range of indexes changes in production. The magnitude of the relative bias for the bootstrap method is about twice of the other methods for P08. The bootstrap also has the largest relative bias value for P07. The TAYLOR method has the largest relative bias for P74. However, the TAYLOR method has the smallest bias for P08. BRR underestimated all three strata, but the relative biases for P07 and P74 were negligible.

The bootstrap method underestimated the standard errors for about half of the strata, while the other methods tended to overestimate the standard errors. The bootstrap is the only method with a negative average relative bias over all of the strata (see Table 1). On the other hand, the bootstrap method has the smallest average relative bias value over all of the strata with or without P07, P08. The TAYLOR and jackknife methods both have large over and underestimated cases. Figure 2 shows the average stability for each stratum. Strata P07 and P08 have the most unstable estimates for all methods. The bootstrap is the most stable method for P07 and P08 with values of about 30% and 14%, respectively. The standard errors for the jackknife and TAYLOR methods

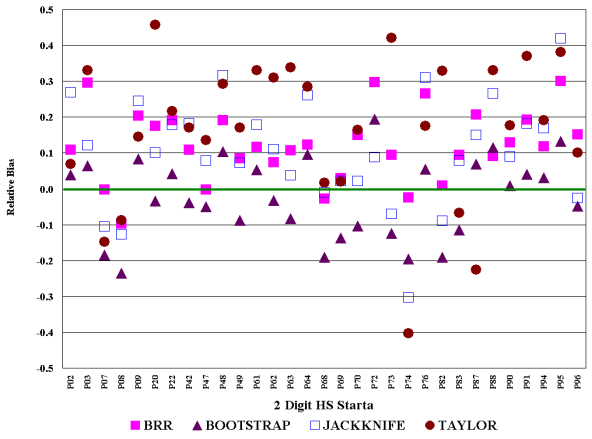

Figure 1: Average Relative Bias of the Standard Errors for the Annual Changes

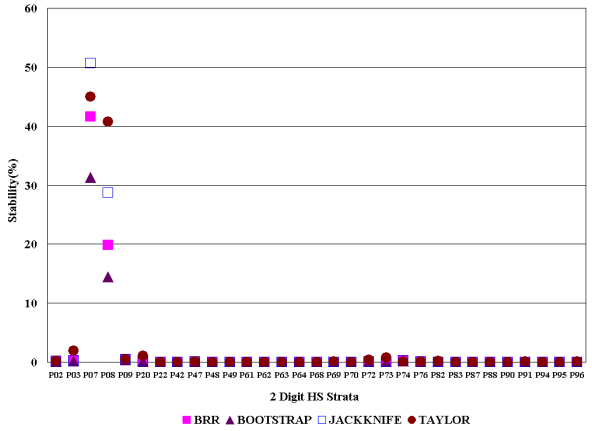

Figure 2: Percentage of Stability for Average of Annual Changes

are the more unstable than the other methods for P07 and P08. There are two other strata with stability measures greater than 1% for the TAYLOR method. All of the other strata and methods have stability measures less than 1%.

Figure 3 shows the same figure as Figure 2 but without P07 and P08. The bootstrap method demonstrated the most stability and showed less variation over all strata than the other methods, except P74. The TAYLOR method demonstrated the most stability for P74 but the most instability over the other strata.

5.3.2 Coverage Rate of the 95% Confidence Interval

Approximately half of the strata with average coverage rates greater than 95% for the bootstrap. However, the bootstrap has only three strata with coverage rates less than 90% and P08 is the only stratum with a coverage rate less than 80%. The TAYLOR method has average coverage rates greater than 95% for about 2/3 of the strata, but also has four strata with coverage rates less

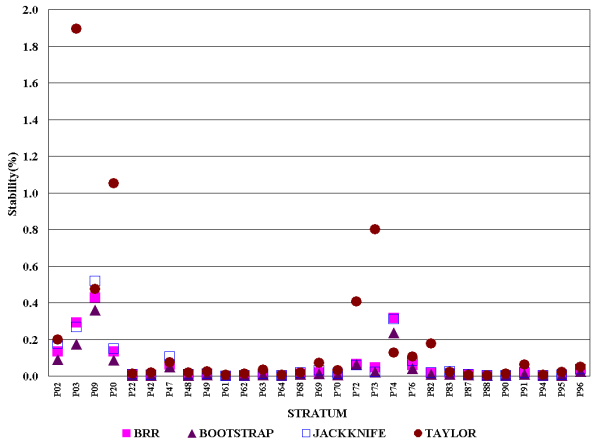

Figure 3: Percentage of Stability for Average of Annual Changes, Without P07, P08

than 80%.

BRR has the highest average coverage rate for P07 and P08, but the average coverage rate for P07 is less than 90% and for P08 the coverage rate is near 80%. These two strata are the only strata with coverage rates less than 94% for BRR. There are 25 strata with average coverage rates of $95 \pm 3\%$ for BRR and bootstrap method. However, the average coverage rates outside $95 \pm 3\%$ range are all above higher than 98% with the exception of P07 and P08. Meanwhile, for the bootstrap the average coverage rates outside the $95 \pm 3\%$ range are all less than 92% except P88. The jackknife and TAYLOR methods are in between with some over coverage and some under coverage. The jackknife and TAYLOR methods both have strata with coverage rates less than 80% even if P07 and P08 are excluded.

BRR is the only method that has an average coverage rate over the strata greater than 95%. The average coverage rate over the strata for the other three methods are similar with or without P07 and P08. The average coverage rates for these three methods are near 95% when P07 and P08 are excluded. The average coverage rate over all the strata increased about 1% when these two strata are excluded.

6. Findings

The bootstrap method underestimated the variance for half of the strata while the other methods tended to overestimate the variance. However, the range of the relative biases for the bootstrap method was similar to BRR and smaller than the Jackknife method and the TAYLOR method. Also, the bootstrap method generally demonstrated the best stability among the methods while the TAYLOR method tended to be the most unstable. The bootstrap method had the smallest average relative bias value among the strata of all methods with small standard deviation.

All of the methods underestimated the standard errors

for P07 and P08. These two strata were also the most unstable strata for each method. Therefore, the coverage rates for the 95% confidence intervals were poor for these strata.

With exclusion of strata P07 and P08, the 95% confidence interval coverage rates were all least 94.5% and the average of the 95% confidence interval coverage rate was higher than 96% for BRR. The bootstrap method had the fewest strata with coverage rates greater than 95% among the methods. However, the bootstrap had the most strata with coverage rates between $95 \pm 2\%$. Moreover, the bootstrap had fewer strata with coverage rate less than 90% than either the jackknife or TAYLOR method.

7. Future Work

Future work for this project includes taking a look at the effect of imputation on the variance estimates, calculating variance estimates for the secondary classification systems, and testing if the results on production data are similar to the results of our simulation study. Estimating the index and variance using sampling weights instead of using fixed index aggregation weights for the classification group and above is another topic of interest for further research.

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

REFERENCES

Bobbitt, P., Cho, M. J. and Eddy, R. M. (2005), "Weighting Scheme Comparison in the International Price Program", *2005 Proceedings of the American Statistical Association. Government Statistics Section*, [CD-ROM], 1006-1014.

Bobbitt, P. A., Paben, S. P., Cho, M. J., Chen, T-C., Himelein, J.A. and Ernst, L. R. (2007) "Application of the Bootstrap Method in the International Price Program", *The Joint Statistics Meeting.*

Brockwell, P. J. and Davis, R.A. (2002) *Introduction to Time Series and Forecasting*, New York, NY: Springer.

Cassella and Bergers, *Statistical Inference.*

Cho, M.J., Chen, T-C., Bobbitt, P.A., Himelein J.A., Paben, S.P., Ernst, L.R. and Eltinge, J.L.(2007), "Comparison of Simulation Methods using Historical Data in the U.S. International Price Program", *The Third International Conference of Establishment Survey.*

Cochran, William G. (1971), *Sampling Techniques, 3rd Edition* John Wiley & Son, Inc.

Diewert, W.E and Feenstra, R.C. (1999) "International trade price indexes and seasonal commodities", *Research paper*, Chapter 7, 4-10.

Himelein, J. A. (2006), "Alternative Approach to POP Taylor Series Variance Methodology", BLS Memorandum dated July 20, 2006.

Hoaglin, Mosteller, and Tukey (1985), *Exploring Data Tables, Trends and Shapes*, John Wiley and Sons, New York.

Judkins, D. R.(1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics*, **6**, 223-239.

Krewski, D. and Rao, J.N.K.(1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods", *Annals of Statistics*, **9(5)**, 1010-1019.

Lahiri, P.(2003), *Statistical Science*, **Vol. 19**, No. 2, 199-210.

Rao, Wu, and Yue(1992), "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology.*

Tukey, J. W.(1977), "Exploratory Data Analysis", *Reading, MA: Addison-Wesley.*

Wang, X. and Himelein J. A.(2005) "A Top-Down Approach of Modeling IPP short-term Relatives".

Wolter, K.M.(1985), *Introduction to Variance Estimation.* Springer-Verlag, New York.