ASSESSING THE IMPACT OF IMPUTATION ON THE SAMPLING VARIANCE OF THE U.S. CONSUMER PRICE INDEX

Sylvia G. Leaver and William E. Larson
U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 3655, Washington, D.C. 20212
Leaver_S@bls.gov

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

Key Words: Stratified random groups; Complex multi-stage design

Imputation of basic-level price change, due primarily to sample attrition and item unavailability, occurs frequently in Consumer Price Index (CPI) production. Though current variance estimation procedures do take basic-level imputations into account, it is of interest to assess the contribution to total variance that these imputations represent.

In this paper we present estimates of the sampling variance of price change for the U. S. Consumer Price Index which exclude the additional component of variation attributable to basic-level imputation. Variances computed using a random groups methodology are given for 1-, 2-, 6-, and 12-month price change lags at the U.S. level. Estimates are contrasted with production variance estimates for the same series.

In Section 1 the official CPI and the elementary geometric price index estimators are described. Section 2 discusses full sample and replicate previous period price imputation. Section 3 presents the construction of random group replicates and gives the production sampling variance estimator for the CPI. Section 4 contrasts the production CPI variance estimates in which missing prices are imputed independently for each replicate with those in which this step is not taken. Conclusions are given in Section 5.

1. Background

For a full discussion of the CPI the reader is referred to Chapter 17 of the *BLS Handbook of Methods*, (1997), and Leaver and Valliant (1995). However, we will describe certain features of the CPI pertinent to this study. The CPI is calculated monthly for the total US metropolitan and urban non-metropolitan population for all consumer items, and it is also estimated at other levels defined by geographic area and item groups such as food, shelter, and apparel.

Prices for the CPI are collected in 87 primary sampling units (PSUs) in 83 geographic areas. Of these PSUs, 31 are self-representing. The remaining 56 were selected according to a stratified design in which one PSU was selected from each of several strata within each of seven index areas, defined as medium to small-sized Metropolitan Statistical Areas (MSAs) for four Census regions (Northeast, Midwest, South, and West) and urban, non-MSAs in three regions (Midwest, South and West.)

For purposes of variance estimation, the sample for each index area is segmented into two or more subsets, called replicate panels. For self-representing PSUs, these replicates are subsets of the sample for the PSU and are selected and constructed independently of each other. For non-self-representing index areas, each replicate comprises the sample for one or two paired PSUs. The number of replicate panels constructed for each index area varies from 2 for most self-representing index areas (New York, Chicago and Los Angeles each have 4) to 12 for the medium sized cities in the South.

The CPI is estimated for consumer commodities and services, grouped into 211 strata for each index area, although not all such indexes are published every month. It is constructed in two stages. In the first or elementary level stage, the price index for an item-area is updated every one or two months via a function of sample price changes called a price relative. Let X_{ia}^{t} denote the index at time *t*, in item stratum *i*, area *a*, relative to time period *0*. Then

$$X_{ia}^{t} = R_{ia}^{t,t-1} X_{ia}^{t-1}$$

where $R_{ia}^{t,t-1}$ denotes the price relative between times *t* and *t*-1 . Since 1999, elementary indexes for most commodities and services are computed using a weighted geometric average (BLS, 1997):

$$R_{ia}^{t,t-1} = \prod_{j \in S_{ia}} \left(\frac{P_{iaj,t}}{P_{iaj,t-1}} \right)^{w'_{iaj}} = e^{\sum_{j \in S_{ia}} w'_{iaj} \ln\left(P_{iaj,t} / P_{iaj,t-1} \right)};$$

those for shelter and the few remaining item strata use a modified Laspeyres formula. Here S_{ia} represents the sample for item *i* in area *a*, *P* represents the price and w' represents the quote-level sampling weight of sample item *j*, normalized to the same sample rotation base for all quotes in the item-index area.

In the second stage of construction, the index for each higher level item *I* and area *A* grouping is computed as a Laspeyres-type weighted sum of elementary indexes:

(1) $X_{IA}^{t} = \sum_{i \in I} \sum_{a \in A} RI_{ia} X_{ia}^{t}$, where

RI_{ia} is the item-area expenditure-based relative importance, computed from the Consumer Expenditure Survey. In addition to the full sample index series, replicate index series are constructed in an analogous manner, using the corresponding sample panel for each item and item aggregate in each index area, and replicate expenditure weights..

Earlier work in estimating the sampling variance of the CPI was largely devoted to the Laspeyres estimator. Dippo and Wolter (1983) compared Taylor series approximations to jackknifing. In a series of papers, Leaver (1990), Leaver et al. (1991), and Leaver and Swanson (1992), a hybrid random-groups-Taylor series approach was used to estimate the sampling variance of the CPI. Leaver and Valliant (1995) compare this hybrid estimator with a stratified random groups estimator using VPLX (Fay, 1998) software. Current official CPI variance estimates are also based on a stratified random groups estimator (Swanson, 1999 and BLS, 2000). Baskin and Leaver (1996) explored variance estimation for the basic geometric means estimator for the housing component of the CPI and Leaver and Cage (1997) investigated sampling variance behavior for a series of alternatively aggregated price indexes using a stratified jackknife method, and Leaver and Larson (2001) investigated sampling variance behavior of a scanner-based experimental index using stratified jackknife methods. This paper builds on these previous studies and is the first to provide estimates of the effect of imputation on the sampling variability for price change for the production CPI estimates.

2. Full Sample vs. Replicate Imputation

In CPI price relative estimation, missing previous period prices for quotes for which current prices are available are imputed by multiplying a collected or imputed price in the previous period by the full sample *t-2* to *t-1* price relative for the item-area. This procedure is repeated to also impute a replicate level price, to be used in replicate relative computation, using the replicate sample *t-2* to *t-1* price relative for the item-area. Separate replicate level price imputation allows for the inclusion of imputation variance in the resulting sampling variance estimates.

In this paper we investigate the effect of using full sample index area relatives (Method 1) versus replicate-level relatives (Method 2) to impute missing *t-1* prices in replicate index computation. The focus of our study were the eight item strata comprising fresh fruits and vegetables. The overall rate of previous price imputation in CPI relative estimation for these item strata averaged about 5.2 percent per month over the study period, though such rates varied by month and item.

3. Variance Estimation Methodology

A random groups variance estimator was used to compute the sampling variability of price change for the eight item strata and their aggregate index using both methods of replicate price imputation. Replicate index series $\{ X_{\{mr\}}^{t} \}$ were constructed for each month t = January 2000,…,December 2001 for each of n_m replicates in each index area m in the following manner: For each replicate series indexed by *mr*, the index was computed by replacing the full sample estimate for area m with its corresponding estimate from replicate panel r and aggregating this with full sample estimates for all other index areas:

$$X_{\{mr\}}^{t} = \sum_{m' \neq m} RI_{m'} \prod_{s=0001}^{t} R_{m'}^{s,s-1} + RI_{m} \prod_{s=0001}^{t} R_{\{mr\}}^{s,s-1}$$

where $R_{\{mr\}}^{t,t-1} = \left(\prod_{q \in mr} \left(\frac{p_q^t}{p_q^{(t-1)}} \right)^{S_{qr}} \right)$ We note that $R_{\{mr\}}^{t,t-1}$ is computed like $R_m^{t,t-1}$ with the exceptions that it uses just the prices $\{ p_q^t, \ p_q^{(t-1)} \}$, assigned to replicate r and that the expenditure share weight, S_{qr} is ratio-adjusted sum to 1 over the quotes assigned to replicate in that area.

$$S_{qr} = \frac{S_q}{\Sigma_{q' \in r} S_{q'}}$$

Replicate estimates of *k*-month price change were then computed as ratios of the relevant replicate indexes: $PC_{\{mr\}}^{t,t-k} = [(X_{\{mr\}}^{t} / X_{\{mr\}}^{t-k}) - 1] * 100.$ The random groups estimator of the variance of $PC^{t,t-k}$ was then:

$$V\left(PC^{t,t-k}\right) = \sum_{m=1}^{38} \sum_{r=1}^{n_m} \frac{1}{n_m(n_m-1)} \left(PC_{\{mr\}}^{t,t-k} - PC^{t,t-k}\right)^2$$

4. Findings

Table 1 below gives estimates of 1-month price change and 1-month price change standard error for each method of quote level imputation for the aggregate of the eight fresh fruits and vegetables strata. For Method 1, full sample imputed prices were used in both full sample and replicate price relative computation, and for Method 2, replicate imputed prices were used. The table also presents the percentage of the quotes for which previous period prices were imputed, and the contribution to the aggregate variance attributable to imputation for the national index series for the item

Table 1. 1-Month Price Change and Standard Errors, for Methods 1 and 2 Price Imputation for Fresh Fruits and Vegetables and Citrus Fruits, U.S. CPI, February 2000-December 2001

Month	Fresh Fruits & Vegetables, 1 Mo PC	Method 1 1 Mo PC Standard Error	Method 2 1 Mo PC Standard Error	Difference in Standard Errors	% of Useable Quotes with t-1 Price Imputed	% of Variance Due to Imputation	Citrus Fruits, 1 Mo PC	Method 1 1 Mo PC Standard Error	Method 2 1 Mo PC Standard Error	% of Useable Quotes with t-1 Price Imputed	% of Variance Due to Imputation
200002	-4.0464	0.6082	0.6247	0.0166	8.55%	5.2375%	2.5997	1.2915	1.3541	7.73%	9.03%
200003	-0.5657	0.5724	0.5912	0.0188	7.18%	6.2608%	-1.5219	1.2951	1.3174	6.61%	3.35%
200004	0.1270	0.4966	0.5260	0.0294	5.25%	10.8721%	-0.9856	1.1769	1.2151	5.03%	6.20%
200005	0.2854	0.6829	0.6794	-0.0035	4.15%	-	2.1691	1.2826	1.3109	4.68%	4.27%
200006	-4.2001	0.5957	0.6136	0.0179	5.82%	5.7418%	0.5036	1.2817	1.3094	8.54%	4.18%
200007	0.3297	0.7345	0.7578	0.0233	7.24%	6.0487%	7.7007	1.5396	1.6201	11.81%	9.70%
200008	0.8492	0.6442	0.6502	0.0060	5.81%	1.8386%	8.3457	1.7428	1.8661	13.13%	12.77%
200009	2.9940	0.6367	0.6646	0.0279	6.75%	8.2329%	2.8223	1.8795	2.0611	13.78%	16.84%
200010	1.5036	0.6340	0.6625	0.0285	8.65%	8.4082%	-5.3816	1.5321	1.7297	21.11%	21.55%
200011	3.3937	0.7887	0.7837	-0.0049	7.24%	-	-16.0799	1.5666	1.5937	19.85%	3.37%
200012	4.6736	0.8291	0.8155	-0.0136	5.70%	-	-4.0344	1.8826	1.8009	16.05%	-
200101	-3.7679	0.7786	0.8021	0.0235	5.95%	5.7857%	1.2093	1.4979	1.5619	9.34%	8.02%
200102	-1.5549	0.8058	0.8483	0.0425	6.83%	9.7642%	2.7547	1.9009	1.9531	6.73%	5.27%
200103	0.4069	0.6269	0.6841	0.0571	6.60%	16.0068%	0.8964	1.4931	1.9391	5.62%	40.71%
200104	2.1440	0.6982	0.7050	0.0068	5.39%	1.9216%	8.9648	1.7974	1.8757	7.00%	8.17%
200105	-0.2749	0.6831	0.7163	0.0332	5.21%	9.0451%	1.2031	0.9383	0.9908	6.22%	10.32%
200106	-0.7795	0.9472	0.9486	0.0014	4.51%	0.2902%	6.8081	1.6333	1.6453	10.05%	1.45%
200107	-1.7051	0.5871	0.6017	0.0146	9.23%	4.7958%	5.6194	1.8240	1.9311	12.33%	10.79%
200108	-1.9350	0.6481	0.6553	0.0072	7.42%	2.1838%	1.4292	1.9189	1.9416	15.39%	2.33%
200109	2.8850	0.6961	0.7292	0.0331	6.84%	8.8724%	5.2238	2.0157	2.0533	16.39%	3.64%
200110	1.6980	0.6563	0.6804	0.0242	9.45%	6.9727%	-5.0634	1.8392	2.1399	20.44%	26.13%
200111	1.4983	0.8232	0.7721	-0.0511	10.32%	-	-10.6518	1.5699	1.6235	25.23%	6.49%
200112	0.8659	0.8296	0.8020	-0.0276	8.29%	-	-8.0961	1.7321	1.7483	23.30%	1.84%

Table 2. 2-, 6-, and 12-Month Price Change Standard Errors, for Methods 1 and 2 Price Imputation for Fresh Fruits and Vegetables and Citrus Fruits, U.S. CPI, January 2001-December 2001

Month	Fresh Fruits & Vegetables, Method 1 2 Mo PCSE	Ratios of 2-Mo Variances Method 1 to Method 2	Fresh Fruits & Vegetables, Method 1 6 Mo PCSE	Ratios of 6-Mo Variances Method 1 to Method 2	Fresh Fruits & Vegetables, Method 1 12 Mo PCSE	Ratios of 12 Mo Variances Method 1 to Method 2	Citrus Fruits, Method 1 12 Mo PC	Method 1 12 Mo PC Standard Error	Ratios of 12-Mo Variances, Method 1 to Method 2
200101	0.7332	1.2584	1.1890	1.1216	1.196012	1.2343	-4.9437	3.2119	1.2243
200102	0.8106	0.9361	1.1403	1.0769	1.157165	1.1648	-4.8002	3.1600	1.3075
200103	0.8008	1.0011	0.8978	1.2382	0.998034	1.1380	-2.4623	3.1240	1.3748
200104	0.7448	0.9135	0.9467	1.1315	1.073804	1.1329	7.3397	3.5096	1.2731
200105	0.8338	0.9887	0.9114	1.0029	1.014178	1.2204	6.3247	3.1159	1.3031
200106	0.9035	0.9542	0.9776	1.0542	1.241815	1.2246	12.9944	3.6266	1.1207
200107	0.9742	1.0260	1.0367	1.0885	1.160661	1.2956	10.8108	4.2396	1.3572
200108	0.6525	1.0291	1.0323	0.9768	1.160267	1.1550	3.7369	4.3850	1.1565
200109	0.6899	0.9779	1.1482	1.0102	1.292487	1.2397	6.1597	4.5939	1.4113
200110	0.7236	0.9606	1.0151	1.0764	1.196774	1.5056	6.5167	3.9670	1.6039
200111	1.1198	1.0665	1.2874	1.1384	1.133489	1.3471	13.4064	3.6307	1.8716
200112	1.3096	1.1381	1.4601	1.1531	1.527864	1.2701	8.6066	2.8182	1.6413

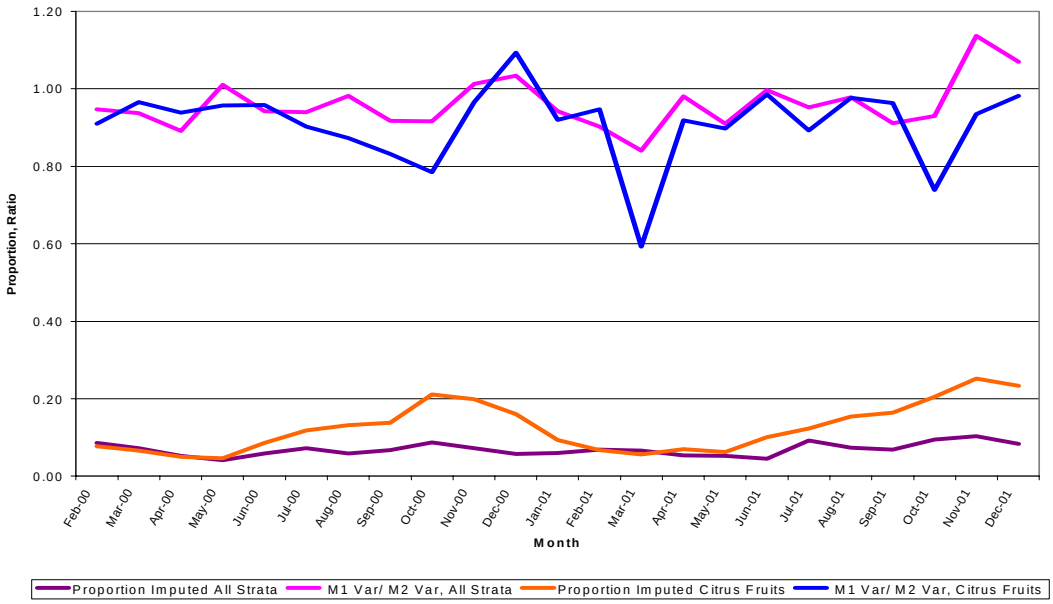
aggregate over the 23-month interval of the study. It also includes the same summary for item stratum FK03, Citrus fruits, which exhibited a somewhat larger and more seasonal quote level imputation rate. From the table we can see that for 1-month price change at the area aggregate level, and particulary for FK03, the contribution of imputation to total variance varies from month to month and tends to track the degree of previous month price imputation that occurs. This effect is not surprising.

What was surprising, and indeed puzzling, was that this effect did not persist with longer intervals. Table 2 gives estimates of Method 1 2-, 6-, and 12-month price change standard error for the 8-item aggregate, and separately for Citrus fruits for 12-month price change for the last 12 months of the study. Also given are the ratios of Method 1 and Method 2 variances, which are shown for the item aggregate in Figure 2. We see that for both the individual item and item aggregate, Method 1 12-month variances are 13% to nearly 90% higher. Examining 2- and 6-month price change estimates, we see that this effect begins to exhibit itself with even 2-month lag estimates, and becomes more pronounced as the length of the lag increases. This result seemed counter-intuitive, as the 1-month price change variances with Method 1 were generally consistently lower. This finding is currently being investigated. Preliminary analyses of Taylor linearized variance estimates, which use weighted index variances and covariances, have been inconclusive, however.

5. Conclusions

The effect of previous price imputation on the sampling variance of short term index change in the CPI appears to directly depend on the degree of imputation that occurs. It appears highest, but at the same time, most variable, in months in which there is significant imputation, such as when seasonal items return to availability. In discussing this phenomenon, one must keep in mind the low level of reliability in variance estimates produced at an index area with typically 1 (with 2 replicates) or a few more degrees of freedom. That this effect does not obtain with longer price change intervals is surprising, as it was expected that national aggregates of 38 index areas would have greater stability and that the effect observed in 1-month change would persist with chaining over successive months. Nevertheless, we are examining estimates for item strata, replicate samples for which one can have few observations. This remains a topic of active investigation.

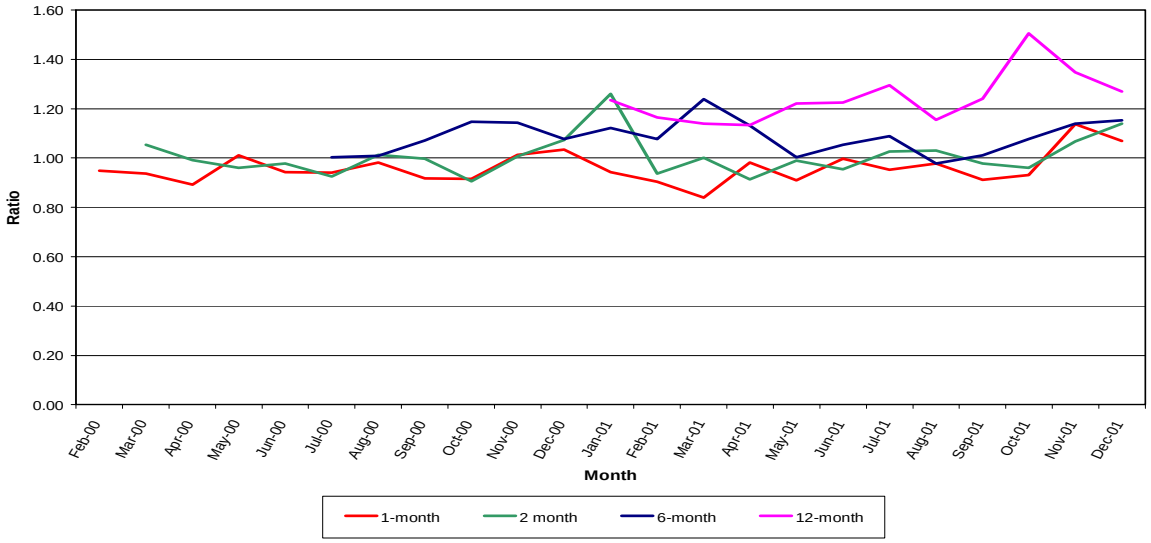
Figure 1. Proportion of Imputed t-1 Prices and Ratio, Method 1 to Method 2 1-Month Price Change Variance Estimates, Fresh Fruits and Vegetables and Citrus Fruits, U.S. CPI, February 2000-December 2001



6. Acknowledgments

The authors would like to thank Janet Williams and Alan Dorfman for their careful reading of this paper. The authors also thank Bill Cook and Gayle Brooks for

Figure 2. Method 1/Method2 Ratios of Price Change Variance, U.S. CPI Fresh Fruits and Vegetables, February 2000-December 2001



their assistance in accessing and processing the CPI microdata. The authors thank Marvin Kasper and Janet Williams for their support on this project.

7. References

Baskin, R.M. and Leaver, S.G. (1996), "Estimating the Sampling Variance for Alternative Forms of the U.S. Consumer Price Index," *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 192-197.

Bureau of Labor Statistics (1997), *BLS Handbook of Methods*, Washington. DC, http://www.bls.gov/opub/hom/homch17_a.htm.

Bureau of Labor Statistics (1997), "The Experimental CPI using Geometric Means (CPI-U-XG)," April 10, 1997 (Washington: Bureau of Labor Statistics).

Bureau of Labor Statistics (2000), "Variance Estimates for Changes in the Consumer Price Index, January 1999-December 1999," *CPI Detailed Report, November 2000*, Washington. DC: U.S. Government Printing Office, pp. 4-6K.

Dippo, C. S., and Wolter, K. M. (1983), "A Comparison of Variance Estimators Using the Taylor Series Approximation, " *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 113-119.

Fay, R. E. (1998), "VPLX, Program Documentation, Vol. 1 Introduction, " U.S. Bureau of the Census, http://www.census.gov/sdms/vplx/

Leaver, S. G., (1990) "Estimating Variances for the U.S. Consumer Price Index for 1978-1986," *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 290-295.

Leaver, S. G., and Cage, R. A. (1997), "Estimating the Sampling Variance for Alternative Estimators of the U.S. Consumer Price Index," *Proceedings of the Survey Research Methods Section, American Statistical Association* pp. 740-745.

Leaver, S. G., Johnstone, J. E., and Archer, K. P. (1991), "Estimating Unconditional Variances for the U.S. Consumer Price Index for 1978-1986," *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 614-619.

Leaver, S. G., and Larson, W. E. (2001), "Estimating Variances for a Scanner-Based Consumer Price Index," *Proceedings of the Government Statistics Section*, American Statistical Association, pp..

Leaver, S. G., and Swanson, D. C. (1992), "Estimating Variances for the U.S. Consumer Price Index for 1987-1991, "*Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 740-745.

Leaver, S. G. and Valliant, R. L. (1995), "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods.* Wiley & Sons, Inc., pp. 543-566.

Swanson, D. C. (1999), "Variance Estimates for Changes in the Consumer Price Index, January 1998-December 1998," *CPI Detailed Report, December 1999*, Washington. DC: U.S. Government Printing Office, pp. 7-20.