ESTIMATING VARIANCES FOR A SCANNER-BASED 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 jackknife; Scanner data

This paper presents estimates of the sampling variance of price change for an experimental scanner-based Consumer Price Index for cereal for the New York Consolidated Metropolitan Statistical Area. Variances are presented for 1-, 2-, 6-, and 12-month price change lags, using a stratified jackknife methodology.

In section one the official CPI and scanner-based geometric price index estimators are described. Section two presents the construction of jackknifed replicates and gives the sampling variance estimator for the scanner index series. Section three presents computational results and contrasts scanner variances with variance estimates for the published price index for cereal for the same area. Sources of price change variability are identified and discussed. Section four contrasts variance estimates in which missing prices are imputed independently for each jackknifed replicate with those in which this step is not taken. Conclusions are given in section five.

1. Publication and Scanner-based Indexes

For a full discussion of the CPI the reader is referred to Chapter 19 of the *BLS Handbook of Methods*, (1992), 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 7 index areas, defined as medium to small-sized Metropolitan Statistical Areas (MSAs) for 4 Census regions (Northeast, Midwest, South, and West) and urban, non-MSAs in 3 regions (Midwest, South and West.)

The CPI is estimated for items 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 1 or 2 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):

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

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.

The index for higher level item *I* and area *A* groupings is computed as a Laspeyres-type weighted sum of elementary indexes:

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

r_{ia}^{b} is the item-area relative importance or relative expenditure share, computed from the Consumer Expenditure Survey for reference period *b*.

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. This paper builds on these previous studies and is the first to provide standard error estimates for a scanner-based index series.

The CPI program office has purchased from A.C. Nielsen Corporation scanner data for cereal from their sample of retail establishments in the New York Consolidated Metropolitan Statistical Area (A101), comprising three PSUs (New York City, A109; New York-Connecticut suburbs, A110; and New Jersey - Pennsylvania suburbs, A111). These data cover all cereal sales, coded at the Universal Product Code (UPC) level, in the Nielsen sample in the New York area, recorded for February 1998 through 2001.

The Nielsen sample is stratified by major chain, and sample stores within chains were selected using a Peano key equal probability selection scheme (Garrett and Harter, 1995). The average sampling rate within a chain is approximately one in ten, but this rate varies by chain. Weekly per unit prices and quantity data are recorded for each UPC in each sample store in which sales occur. Quantity values for each sample store-UPC are inflated by a projection factor, which is the ratio of the store's total sales to chain-level total sales for the same week. Total cereal sales estimates are computed by multiplying reported per unit prices by projected quantities.

Using these data, the CPI program office has constructed a series of alternative scanner-based price indexes for cereal (Richardson, 2000). This study examines the one of these index series which is most similar in its construction to the official CPI.

Like the official CPI, the scanner-based index is constructed in two stages. In the first or elementary stage, the index for each area m is computed as the product of month-to-month price relatives:

$X_m^t = \prod_{s=9803}^{t} R_m^{s,s-1}$. The price relative considered in this study is a chain-based geometrically averaged scanner price relative:

$R_m^{t,t-1} = \prod_{c=1}^{n_m} \prod_{\substack{q \in c \\ c \in m}} I_q^{t,t-1} \left(\frac{p_q^t}{p_q^{t-1}} \right)^{S_q}$. Here $I_q^{t,t-1}$ is the indicator function for the sales for product q (usually an item corresponding to a unique UPC or two or more UPCs judged to be sufficiently similar to combine) in chain c at both times t and t - 1, and S_q is the expenditure share of q, i.e., the ratio of the previous year's expenditure for q to the sum of the previous year's expenditures for all items available at both times t and t - 1. Counts n_m and n_{mc} refer to the number of chains in index area m and the number of products sold in chain c in index area m, respectively. The price p_q^t is computed as the unit value price per ounce of product q in chain c at time t,

$p_q^t = \frac{\sum_{i=1}^{n_{mc}} P_{iq}^t Q_{iq}^t}{\sum_{i=1}^{n_{mc}} Q_{iq}^t \; Z_{iq}^t}$, where P_{iq}^t, Q_{iq}^t and Z_{iq}^t are the price, quantity in units, and size in ounces per unit, respectively, of q in store i at time t. Sales data for the first three weeks in each month are averaged to produce unit valued prices.

Aggregation of scanner indexes proceeds as described in formula (1) above and estimates of k-month percentage price change relative to time t are obtained by taking the ratio of the index at time t to its value at time t-k:

$PC_{A101}^{t,t-k} = [(X_{A101}^{\,t} \big/ X_{A101}^{t-k}) - 1]*100.$

2. Variance Estimation Methodology

A stratified jackknife variance estimator was used to evaluate the sampling variability of price change for the scanner index series. The estimator was based on a segmentation of the Nielsen sample into separate index area-chain-identified strata: one for each of 3-8 major chains and one for the remaining scanner outlets within each of the three New York index areas. Within each of the strata in each index area the sample was further grouped into clusters, with each cluster comprising the sample from one or more of the stores in the chain. A total of 126 clusters were identified for the time period of this study: 26 in A109, 53 in A110, and 47 in A111.

Replicate index series $\{ X_{mcw}^t \}$were then constructed, in a manner analogous to that for the full sample index, for each month t = March 1998,…,October 2000 for each of n_{mc} clusters in each of n_m chain-level strata in each index area m. For each replicate series indexed by mcw, the price relative was computed by deleting the sample for cluster w in stratum c in area m, reweighting the sample for the other clusters in stratum c, and using the full sample for all other strata:

$$X_{mcw}^{\,t} = \sum_{m' \neq m} r_{m'} \prod_{s=9803}^{t} R_{m'}^{s,s-1} + r_m \prod_{s=9803}^{t} R_{mcw}^{s,s-1}$$

where

$$R_{mcw}^{t,t-1} = \left(\prod_{\substack{c'=1 \\ c' \neq c}}^{n_m} \prod_{q \in mc'} I_q^{t,t-1} \left(\frac{p_q^t}{p_q^{t-1}} \right)^{S_q} \right) \left(\prod_{q \in mc} I_q^{t,t-1} \left(\frac{p_{qw}^t}{p_{qw}^{(t-1)}} \right)^{S_{qw}} \right)$$

We note that $R_{mcw}^{t,t-1}$ was computed like $R_m^{t,t-1}$ with the exceptions that for chain c it used the prices p_{qw}^t, and $p_{qw}^{(t-1)}$, the unit values computed with cluster w, stratum *c*, area *m* omitted, and the expenditure share weight, S_{qw} was ratio-adjusted to reflect the loss of expenditure share from the omitted cluster *w*.

$$p_{qw}^{t} = \frac{\sum_{\substack{i=1 \\ i \neq w}}^{n_{mc}} P_{iq}^{t} Q_{iq}^{t}}{\sum_{\substack{i=1 \\ i \neq w}}^{n_{mc}} Q_{iq}^{t} \, Z_{iq}^{t}}, \quad S_{qw} = S_{q} \frac{\sum_{i=1}^{n_{mc}} \sum_{q'} S_{q'}}{\sum_{\substack{i=1 \\ i \neq w}}^{n_{mc}} \sum_{q'} S_{q'}}$$

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

$$V\left(PC^{t,t-k}\right) = \sum_{m=1}^{3} \sum_{c=1}^{n_m} \frac{n_{mc}-1}{n_{mc}} \sum_{w=1}^{n_{mc}} \left(PC_{mcw}^{t,t-k} - PC^{t,t-k}\right)^2$$

3. Findings

Table 1 below gives estimates of average 1-month price change and the contribution to average aggregate variance for each index area and their aggregate for the scanner series over the 33-month interval of the study. From the table we can see that the principal component of variance was consistently from the A109 area, representing, for 1-month change, over 75% of the total sampling variability. Analysis at the stratum level revealed that this component's magnitude derived from the "leftover" stratum, which consists of stores sampled from smaller chains and independent grocers. This stratum represents over 65% of the expenditure weight in A109 for both 1998 and 1999, and exhibited remarkably more price change variability than the remaining three strata in A109.

Table 1. 1-Month Price Change and Sampling Variance for a Scanner-Based Price Index for Cereal, NY A101, March 1998-October 2000, Original Stratification

Area	Avg 1-Mo Price Change (%)	Avg PC Variance, Contribution to Avg 1-Mo PC Variance for A101
Scanner, A101	0.11467	0.22666
A109	0.04756	0.18045
A110	0.13330	0.02094
A111	0.15725	0.02527

We then considered further division of the principal strata, with particular interest in the large stratum in A109. Examining store-level relatives in A109 plotted by stratum, we discovered that while there was considerable coherence of price change between stores within the certainty strata, there was a high degree of variability between stores within the remainder stratum, with price increases and decreases for different stores occurring within the same month. We then computed estimates of price change correlation, which, coupled with additional information from Nielsen revealed that the remainder stratum contained identifiable substrata: two in A109 and three in A110. Figure 1 shows this behavior in A109. The two substrata are shown in darker colors.

Consequently, the identified substrata in that stratum in PSUs A109 and A110 were treated as separate strata, and indexes and variances were recalculated. This restratification achieved a reduction in the New York A101 1-month price change variance of over 50%. Figure 2 depicts the reduction in sampling variance achieved by the additional substratification of A109 and A110.

Table 2 below gives estimates of average 1-, 2-, 6- and 12-month price change and average variance for each index area and their aggregate over the 33-month interval of the study, for the restratified scanner series, and the same for the A101 area for the CPI series as well. Figures 3-4 depict 1- and 12-month price change estimates and their 2-standard error bands for the scanner and official CPI series.

We see from the Figure 3 that, with a few remarkable exceptions, scanner month-to-month price change lies within the 2-standard error bands for the CPI for the same area. The two instances in which this is not the case explain the larger differences that appear between 12-month estimates shown in Figure 4.

The official CPI for cereal is remarkably more variable than the scanner series. This is hardly surprising, given the dramatic differences in available sample size for the two estimators. The CPI for cereal in A101 had, over the period of this study, approximately 55 quotes available for month to month price change estimation, distributed among the three index areas. The Nielsen data set contains approximately 115,000 UPC-store-week-level price-quantity observations for the first three weeks of each month. These collapse down to 6000-6600 monthly chain-UPC group observations. Given the large disparities in sample size, we had expected a greater than six- to seven-fold difference between the two series in their sampling error estimates. This did not happen.

4. 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 good or imputed price in *t*-2 by the full sample *t*-2 to *t*-1 price relative for the item-area. An analogous procedure is applied to the scanner series: full sample index area relatives are used to impute missing *t*-1 unit-valued price data in full scanner price relative computation. The overall rate of imputation in full sample scanner relative estimation is about three percent per month.

We investigated the effect of using replicate-level index area relatives versus full sample relatives to impute missing *t*-1 prices in replicate index computation. Figure 5 displays the percentage difference in standard errors between the two imputation methods for the period of the study. In no case were these differences large; both imputation methods produced stable estimates and the imputation rates are very low. The largest differences occurred most often in A109, where sampling variability was the greatest.

5. Conclusions

The current research indicates that computation of a sampling variance estimate for a scanner-based price index is feasible. Comparison of full sample versus replicate-level imputations yielded very small differences in this application. This was due largely to the extremely low imputation rate for the scanner-based index. Careful restratification of the sample in the large remainder stratum produced a substantial reduction in the sampling error of the resultant price change estimator for the New York CMSA. And, though the sampling variability of the scanner-based cereal index is significantly smaller than that of the CPI, the high variability of cereal price change, particularly among outlets within the heavily weighted and comparatively thinly sampled remainder stratum in A109, is responsible for the much larger than expected estimates of sampling variance for the scanner index.

6. Acknowledgments

The authors would like to thank Janet Williams and Alan Dorfman for their careful reading of this paper. The authors also thank David Richardson and Lyuba Rozental for their assistance in accessing and processing the Nielsen scanner data. 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 (1992), *BLS Handbook of Methods*, Washington. DC: U.S. Government Printing Office, pp. 176-235.

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/

Garrett, J. K. and Harter, R. M. (1995), "Chapter 10: Sample Design using Peano Key Sequencing in Market Research" *Business Survey Methods*, . Wiley & Sons, Inc., pp.201-218.

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 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.

Richardson, D. H. (2000), "Scanner Indexes for the CPI," *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, http://www.nber.org/books/criw00/index.html

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.

Figure 1. 1-Month Price Change for Individual Stores with 2 Substrata, PSU A109, Remainder Stratum, March 1998-October 2000

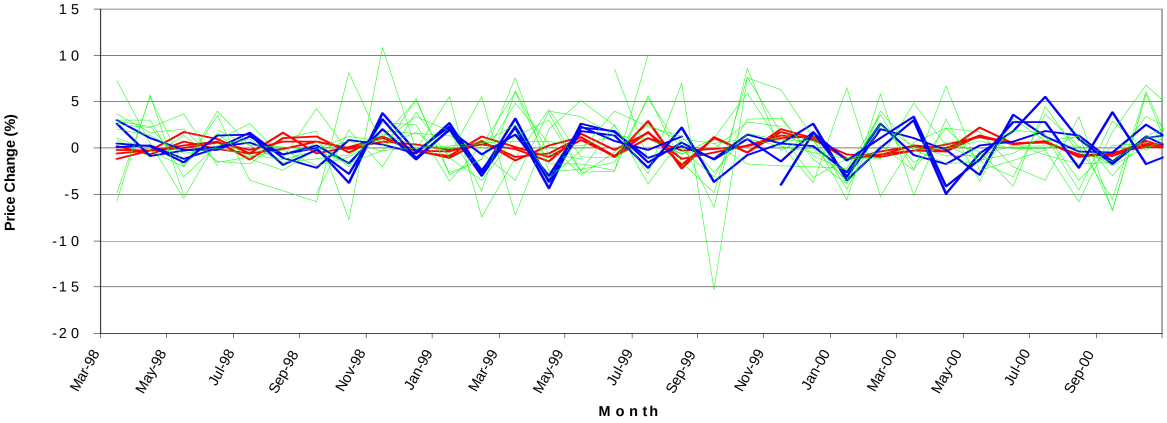

Figure 2. 1-Month price change variance estimates, New York CMSA cereal scanner index, original vs. revised stratification, March 1998-October 2000

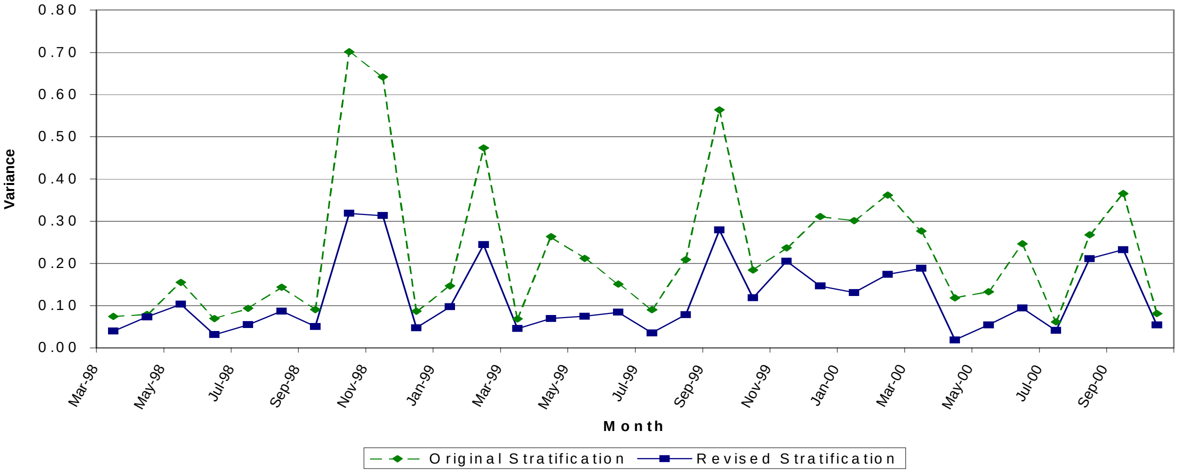

Figure 3. 1-Month price change +/- 2SEs, New York CMSA cereal index CPI vs. scanner, March 1998-October 2000

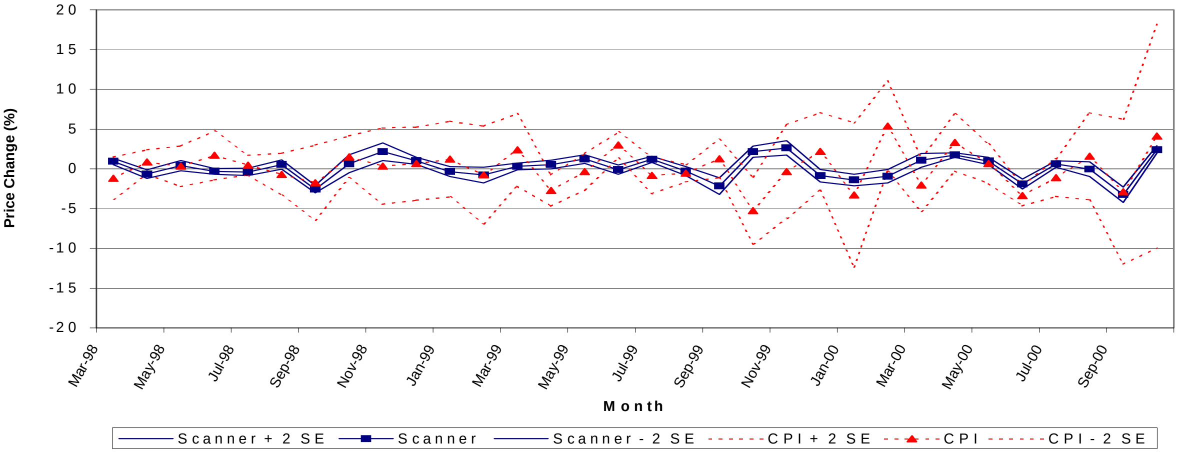

Figure 4. 12-Month price change +/- 2SEs, New York CMSA cereal index
CPI vs. scanner, March 1998-October 2000

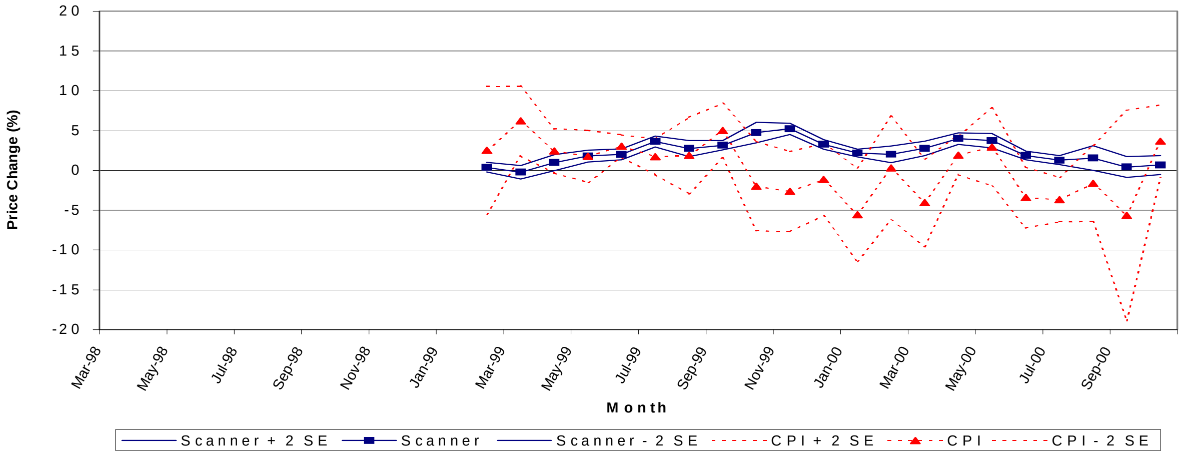

Figure 5. Percentage difference in 1-Month price change standard error
New York CMSA cereal scanner index, replicate vs. full sample imputation
March 1998-October 2000

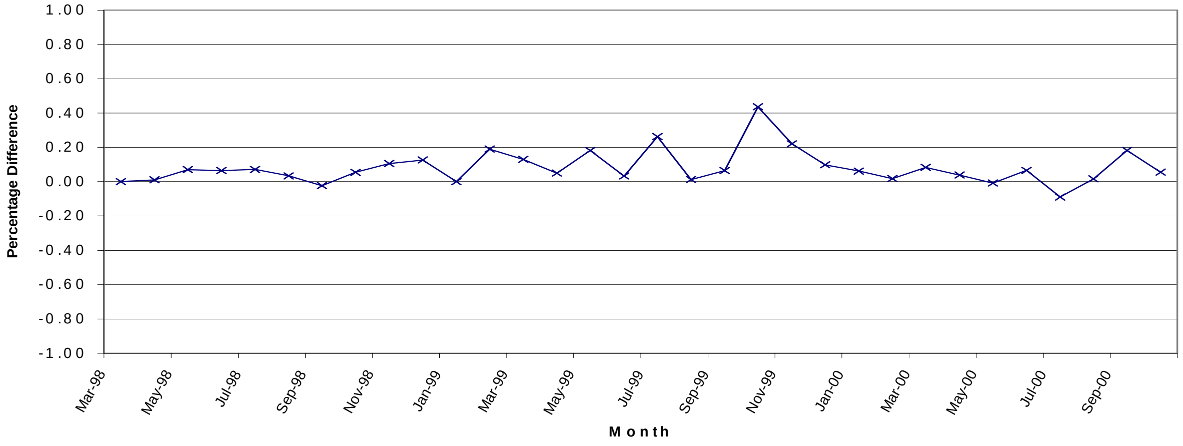

Table 2. 1-, 2-, 6-, and 12-Month Price Change and Sampling Variance for the CPI and a Scanner-Based Price Index for Cereal, New York A101, March 1998-October 2000

Area	Avg 1-Mo Price Change (%)	Avg 1-Mo Price Change Variance, for A101 and PSUs	1-Mo Price Change CV	Avg 2-Mo Price Change (%)	Avg 2-Mo Price Change Variance, for A101 and PSUs	Avg 6-Mo Price Change (%)	Avg 6-Mo Price Change Variance, for A101 and PSUs	Avg 12-Mo Price Change (%)	Avg 12-Mo Price Change Variance, for A101 and PSUs	12-Mo Price Change CV
CPI, A101	0.10807	5.87109	22.42028	0.14278	4.80698	0.52646	6.06046	0.55734	6.71285	4.64875
Scanner, A101	0.12495	0.11839	2.75375	0.14844	0.13299	1.00850	0.15852	1.02276	0.20315	0.44069
A109	0.10365	1.03268	9.80422	0.05976	1.20725	1.00610	1.38919	1.01940	1.74434	1.29560
A110	0.11472	0.07269	2.35014	0.18457	0.10368	1.00878	0.17689	1.02183	0.21927	0.45826
A111	0.15725	0.16749	2.60256	0.19225	0.14638	1.01016	0.16324	1.02612	0.22409	0.46133