Survey Methodology, December 1992
Vol. 18, No. 2, pp. 279–292
Statistics Canada

A Multivariate Procedure Towards Composite Estimation of Consumer Expenditure for the U.S. Consumer Price Index Numbers

P. LAHIRI and WENYU WANG[1]

ABSTRACT

We consider the problem of estimating the "cost weights" and "relative importances" of different item strata for the local market basket areas. The estimation of these parameters is needed to construct the U.S. Consumer Price Index Numbers. We use multivariate models to construct composite estimators which combine information from relevant sources. The mean squared errors (MSE) of the proposed and the existing estimators are estimated using the repeated half samples available from the survey. Based on our numerical results, the proposed estimators seem to be superior to the existing estimators.

KEY WORDS: Consumer expenditure; Composite estimation; Consumer Price Index; Cost weight; Diary survey; Half sample; Laspeyres Index; Mean squared error; Synthetic estimation.

1. INTRODUCTION

The U.S. Consumer Price Index (CPI) is an indicator of price changes for a set of items, goods and services, whose quantity and quality are fixed over a period of time. The U.S. Bureau of Labor Statistics (BLS) computes a number of consumer price indices each month for various geographical areas, consumer units and item classification (*vide* BLS Handbook of Methods 1988).

The smallest group of item classification for which the BLS computes the CPI is known as an "item stratum". It is a prespecified set of consumer goods and services, *e.g.*, fresh whole milk, which can be purchased in the retail market during a "base period" by a specified set of consumer units. A consumer unit may consist of all members of a particular household related by blood, marriage, adoption, or other legal arrangements. A number of item strata constitutes an expenditure class (*e.g.*, dairy products).

The U.S. is divided into eight major areas for sampling purposes. A major area may be either "self-representing" or "non-self-representing" and belongs to one of the four regions (Northeast, Midwest, South and West). A self-representing area consists of all large cities within a region. A non-self-representing area generally consists of a county or a group of contiguous counties. For publication purposes, a major area is further divided into a number of "market basket areas" or "publication areas".

The Laspeyres formula used by the BLS to compute the CPI for a given area and an expenditure class (say, E) is defined below. Let

P_{it} = the average price of all items in the ith item stratum at time t ($t = 0,T$),

Q_{i0} = the quantity of all items in the ith item stratum purchased at time $t = 0$ (base period).

[1] P. Lahiri, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588 0323, USA. Wenyu Wang, SUNY Health Science Center at Brooklyn, Box 1203, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.

Then the Laspeyres index at time $t = T$ is given by

$$I_T = \sum_{i\epsilon E} Q_{i0}P_{iT} \Big/ \sum_{i\epsilon E} Q_{i0}P_{i0}$$

$$= \frac{\sum_{i\epsilon E} C_i(P_{iT}/P_{i0})}{\sum_{i\epsilon E} C_i}$$

$$= \sum\nolimits_{i\epsilon E} R_i(P_{iT}/P_{i0}),$$

where

$C_i = Q_{i0}P_{i0}$ = total expenditure for all items in the ith item stratum at $t = 0$,

$R_i = C_i/\sum_{i\epsilon E} C_i$ = proportion of total expenditure spent on the ith item stratum at $t = 0$.

The quantities C_i and R_i are referred to as the "cost weight" and "relative importance" of the ith item stratum within the expenditure class, E.

The Bureau of Labor Statistics computes the consumer price indices using data from the U.S. Consumer Expenditure Survey (CES). The survey has two different components - Diary survey and Interview survey, each having separate sampling schemes and questionnaires. In this paper we consider data from the Diary survey only. The sampling design selects all the primary stage units (PSU's) within a particular self-representing area with certainty. But only a sample of PSU's is selected for a particular non-self-representing area according to a probability sampling scheme. From each selected PSU, a sample of consumer units (CU's) is selected again using some probability sampling design. Each respondent keeps a diary of expenditures on various items for two consecutive 1-week periods. For a detailed account on the CPI and CES, the reader is referred to the BLS Handbook of Methods (1988).

The efficiency of the traditional sample survey estimators of the cost weight and relative importance of an item stratum at the publication area level is generally very low compared to their efficiency at a larger area (*e.g.*, major area) level. This is due to the fact that only a few consumer units are available from a given publication area. Thus, there is a need to improve the traditional estimator by borrowing strength from related resources. Marks (1978) and Cohen and Sommers (1984) considered certain composite estimators which pool information from related areas. Ghosh and Sohn (1990) obtained composite estimators of the cost weight and relative importance using an empirical Bayes approach.

The current procedure used by the Bureau of Labor Statistics consists of several steps. First composite estimators of the relative importances are obtained using a method suggested by Cohen and Sommers (1984). The estimators of the cost weights are then obtained from these estimators of the relative importances using an iterated "raking" procedure. The final estimates of the cost weights for the entire expenditure class and for the major area are identical to the corresponding preliminary estimates. One reason for ensuring this "data consistency" by raking may be due to the fact that the performances of the preliminary estimators are generally satisfactory at a higher level of aggregation compared to their performances at a lower level. At the last step, the final estimators of the relative importances are obtained directly from the final cost weight estimators by division.

Unlike earlier authors, we use the correlations between the item strata in proposing our composite estimators in Section 2. The shrinkage factor of the composite estimator obtained by minimizing the mean squared error within an appropriate class of estimators involves some unknown parameters. These unknown parameters are estimated using the balanced repeated replications available from the survey. The estimator proposed by Cohen and Sommers (1984) turns out to be a special case of our estimator if one assumes that the preliminary estimators are all uncorrelated.

In Section 2 we concentrate our attention to the estimation of the cost weight of an item stratum for a publication area. However, we can obtain estimators of the cost weights at a higher level of aggregation (*e.g.*, expenditure class for a publication area, *etc.*) by appropriate summation. From our study, it turns out that in terms of the mean squared error criterion these estimators always perform better than the corresponding preliminary estimators and hence better than the BLS estimators (note that due to the raking procedure the BLS estimators are identical to the preliminary estimators at higher levels of aggregation).

In Section 3 we propose a composite estimator of relative importance of an item stratum at the publication area level. Instead of using the preliminary estimators of the cost weights we use the preliminary estimators of the relative importances for all the item strata belonging to the expenditure class under consideration. The preliminary estimators of relative importances of all the item strata within an expenditure class add up to unity. Thus, the variance covariance matrix of the preliminary estimators is singular and this makes the problem different from the problem of estimation of the cost weights. Our procedure deletes one item stratum in an optimal manner and thus avoids the problem of singularity of the variance covariance matrix of the preliminary estimators. Our numerical results show that in terms of the mean squared error criterion the proposed estimator is always the best among all the rival estimators considered.

In Section 4, we present all the numerical results. We have evaluated different estimators of the cost weight and relative importance based on estimated mean squared error obtained by using the balanced repeated half samples (see McCarthy 1969, Ghosh and Sohn 1990). Based on our results, the proposed estimators seem to be superior to all the rival estimators considered in the paper.

2. ESTIMATION OF THE COST WEIGHT

Let X_{ijl} be the average of two consecutive weeks of expenditure for all the items in the ith item stratum by the lth consumer unit belonging to the jth publication area within a particular major area ($i = 1, \ldots, I; j = 1, \ldots, m; l = 1, \ldots, n_j$). Let W_{jl} be the sampling weight attached to the lth consumer unit in the jth publication area ($j = 1, \ldots, m; l = 1, \ldots, n_j$). This represents a number of consumer units in the population and is obtained by the Census Bureau using a complex procedure which takes into account various factors such as inclusion probabilities, nonresponse, *etc.* In this section, we consider estimation of θ_{ij}, the true average weekly expenditure per consumer unit for the ith item stratum and jth publication area. The cost weight is simply defined as $N_j\theta_{ij}$, where N_j denotes the total number of consumer units in the jth publication area. The preliminary estimator of θ_{ij} is given by

$$Y_{ij} = \sum_{l=1}^{n_j} W_{jl}X_{ijl} \Big/ \sum_{l=1}^{n_j} W_{jl}, \; (i = 1, \ldots, I; \;\; j = 1, \ldots, m). \tag{2.1}$$

Similarly, the corresponding estimator for the major area is given by

$$Y_{i\cdot} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl} \Big/ \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}. \tag{2.2}$$

The variability of $Y_{i\cdot}$ is much lower than that of Y_{ij}. Thus, a composite estimator of θ_{ij} which increases the precision is needed. Let $Y_j = (Y_{1j}, \ldots, Y_{Ij})'$ and $\theta_j = (\theta_{ij}, \ldots, \theta_{Ij})'$, $j = 1, \ldots, m$. Let V_j be the true variance covariance matrix of Y_j, $(j = 1, \ldots, m)$. Under a synthetic assumption, *i.e.*, $\theta_j = \mu$, a $I \times 1$ column vector, $(j = 1, \ldots, m)$, the best estimator of θ_j is given by

$$\tilde{\mu} = \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \sum_{j=1}^{m} V_j^{-1} Y_j, \tag{2.3}$$

which is obtained by minimizing $\sum_{j=1}^{m} (Y_j - \mu)' V_j^{-1} (Y_j - \mu)$ with respect to μ. The synthetic assumption, however, is hardly satisfied. In the other extreme when there is absolutely no similarity between the θ_j's, it is appropriate to take Y_j as an estimator of θ_j. When the real situation is in between these two extremes one may take a composite estimator given by

$$\hat{\theta}_{ij}(a_{ij}) = (1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu}, \tag{2.4}$$

where a_{ij}'s are constants $(0 \le a_{ij} \le 1)$, e_i is a $I \times 1$ column vector having 1 for the ith elements and 0 for the others.

We obtain a_{ij} by minimizing the mean squared error

$$E[\{(1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu} - \theta_{ij}\}^2 \mid \theta_{ij}] \tag{2.5}$$

with respect to a_{ij}. The optimal choice is given by

$$\tilde{a}_{ij} = \frac{e_i' \left[V_j - \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \right] e_i}{E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]}. \tag{2.6}$$

Thus, the optimal estimator of θ_{ij} in the class described by (2.4) is given by

$$\tilde{\theta}_{ij} = (1 - \tilde{a}_{ij}) Y_{ij} + \tilde{a}_{ij} e_i' \tilde{\mu}. \tag{2.7}$$

Remark 1: In the derivation of the optimal estimator $\tilde{\theta}_{ij}$, the quantities V_j, $(j = 1, \ldots, m)$ and $E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]$ are assumed to be fixed and known.

Remark 2: The estimator proposed by Cohen and Sommers (1984) can be obtained from $\tilde{\theta}_{ij}$ as a special case when

$$V_j = \left(\sum_{l=1}^{n_j} W_{jl} \right)^{-1} \text{Diag}(\sigma_1^2, \ldots, \sigma_I^2).$$

Note that according to their assumption the correlation between any two item strata is zero which appears to be very restrictive from our study.

Remark 3: Note that using a familiar matrix inversion result (see Rao 1973),

$$V_j - \left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1} = V_j\left[V_j + \left(\sum_{s\neq j} V_s^{-1}\right)^{-1}\right]^{-1} V_j$$

which is positive definite. Also,

$$E[(Y_{ij} - e_i'\tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m] = e_i'V_j\left[V_j + \left(\sum_{s\neq j} V_s^{-1}\right)^{-1}\right]^{-1} V_j e_i$$

$$+ \left[\theta_{ij} - e_i'\left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1}\left(\sum_{j=1}^{m} V_j^{-1}\theta_j\right)\right]^2.$$

Also, when $\theta_j = \mu$, one gets $\tilde{a}_{ij} = 1$ and thus $\tilde{\theta}_{ij} = e_i'\tilde{\mu}$. Otherwise the size of the shrinkage factor depends on the size of

$$\left[\theta_{ij} - e_i'\left(\sum_{j=1}^{m} V_j^{-1}\right)^{-1}\left(\sum_{j=1}^{m} V_j^{-1}\theta_j\right)\right]^2.$$

The larger the distance of θ_{ij} from $e_i'(\sum_{j=1}^{m} V_j^{-1})^{-1}(\sum_{j=1}^{m} V_j^{-1}\theta_j)$ the smaller is the size of $\tilde{a}_{ij}$. This means that if a particular area is very different from the general nature of all the areas then our procedure will give less weight on the synthetic part of the estimator. This explains the great deal of variation of the shrinkage factors in Table 1.

We shall estimate $\tilde{a}_{ij}$ using the 20 balanced repeated half samples available from the survey. Let $w_{jl}^{(k)}$ denote the weight assigned to the lth consumer unit of the jth area for the kth replication ($j = 1, \ldots, m$; $l = 1, \ldots, n_j$; $k = 1, \ldots, 20$). These replicated weights are constructed by the Census Bureau using a complex procedure. For any replication, approximately half the consumer units receive zero weights and the remaining consumer units receive positive weights.

Table 1

Shrinkage Factors $\hat{a}_{ij}$ in West Non-Self-Representing Area

i j	1	2	2
1	0.8479225	0.7057626	0.9214804
2	0.8434894	0.5692695	0.8092725
3	0.0969009	0.0786758	0.6953904
4	0.4446537	0.5444809	1
5	0.6999551	0.3460123	0.5487382
6	0.0318442	0.4981756	0.2598752

Define

$$\hat{a}_{ij}^{*} = \frac{e_i'\left[\hat{V}_j - \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\right]e_i}{\frac{1}{20}\sum_{k=1}^{20} [Y_{ij}^{(k)} - e_i'\hat{\mu}^{(k)}]^2},$$

$$\hat{\mu} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1}Y_j\right],$$

$$\hat{\mu}^{(k)} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1}Y_j^{(k)}\right],$$

$$Y_{ij}^{(k)} = \sum_{l=1}^{n_j} W_{jl}^{(k)}X_{ijl}\Big/\sum_{l=1}^{n_j} W_{jl}^{(k)},$$

$$Y_j^{(k)} = [Y_{1j}^{(k)}, \ldots, Y_{Ij}^{(k)}]',$$

$$\hat{V}_j = 1/20\sum_{k=1}^{20} [Y_j^{(k)} - Y_j][Y_j^{(k)} - Y_j]'.$$

Then we propose the following estimator of θ_{ij}:

$$\hat{\theta}_{ij}^{*} = (1 - \hat{a}_{ij}^{*})Y_{ij} + \hat{a}_{ij}^{*}e_i'\hat{\mu}. \tag{2.8}$$

Remark 4: Using argument given in Remark 3, $\hat{a}_{ij}^{*} \geq 0$. But it is possible that sometimes $\hat{a}_{ij}^{*}$ may exceed unity. Thus, we consider the following estimator:

$$\hat{\theta}_{ij} = (1 - \hat{a}_{ij})Y_{ij} + \hat{a}_{ij}e_i'\hat{\mu}, \tag{2.9}$$

where $\hat{a}_{ij} = \min[1,\hat{a}_{ij}^{*}]$.

In Table 1, we give values of $\hat{a}_{ij}$ for the West non-self-representing area.

3. ESTIMATION OF THE RELATIVE IMPORTANCE

Let $R_{ij} = Y_{ij}/\sum_{i=1}^{I}Y_{ij}$ be the preliminary estimator of the relative importance $r_{ij} = \theta_{ij}/\sum_{i=1}^{I}\theta_{ij}$, $(i = 1, \ldots, I; j = 1, \ldots, m)$. Let $R_j = (R_{1j}, \ldots, R_{Ij})'$, $(j = 1, \ldots, m)$. Since $\sum_{i=1}^{I} R_{ij} = 1$, $(j = 1, \ldots, m)$, the variance covariance matrix of R_j is singular. Thus, the method described in Section 2 is not directly applicable to this situation. In order to avoid this singularity problem, we delete one item stratum from the expenditure class under consideration. Without any loss of generality, let the Ith item stratum be deleted. Then apply the procedure described in Section 2 to obtain the following estimator for r_{ij}, $(i = 1, \ldots, I - 1; j = 1, \ldots, m)$

$$\hat{r}_{ij}^{*} = (1 - \hat{d}_{ij})R_{ij} + \hat{d}_{ij}e_i'\hat{\xi}, \tag{3.1}$$

where

$$\hat{d}_{ij} = \min[1, \hat{d}^*_{ij}],$$

$$\hat{d}^*_{ij} = \frac{e_i'\left[\hat{D}_j - \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\right]e_i}{\frac{1}{20}\sum_{k=1}^{20} [R_{ij}^{(k)} - e_i'\hat{\xi}^{(k)}]^2},$$

$$R_{ij}^{(k)} = Y_{ij}^{(k)} \Big/ \sum_{i=1}^{I} Y_{ij}^{(k)},$$

$$R_j^{(k)} = (R_{1j}^{(k)}, \ldots, R_{I-1j}^{(k)})',$$

$$\hat{D}_j = \frac{1}{20}\sum_{k=1}^{20} (R_j^{(k)} - R_j)(R_j^{(k)} - R_j)',$$

$$\hat{\xi}^{(k)} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{D}_j^{-1}R_j^{(k)}\right],$$

$$\hat{\xi} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{D}_j^{-1}R_j\right].$$

For $i = I$,

$$\hat{D}_{II}^{(j)} = \frac{1}{20}\sum_{k=1}^{20} (\hat{R}_{Ij}^{(k)} - R_{Ij})^2,$$

$$R_{I\cdot} = \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}R_{Ij}\right] \Big/ \sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1},$$

$$\hat{d}_{Ij} = \min[1, \hat{d}^*_{Ij}],$$

$$\hat{d}^*_{Ij} = \frac{\hat{D}_{II}^{(j)} - \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}\right]^{-1}}{\frac{1}{20}\sum_{k=1}^{20} [R_{Ij}^{(k)} - R_{I\cdot}^{(k)}]^2},$$

$$R_{I\cdot}^{(k)} = \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}R_{Ij}^{(k)}\right] \Big/ \sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1}.$$

We estimate r_{Ij} by a univariate procedure which yields the following estimator of r_{Ij}, $(j = 1, \ldots, m)$:

$$\hat{r}^*_{Ij} = (1 - \hat{d}_{Ij})R_{Ij} + \hat{d}_{Ij}R_I.$$

We obtain the final estimator of r_j as $\hat{r}_j = (\hat{r}_{1j}, \ldots, \hat{r}_{Ij})'$, where $\hat{r}_{ij} = \hat{r}^*_{ij}/\sum_{i=1}^{I}\hat{r}^*_{ij}$. There are I possible choices of deleting one item stratum. We choose the combination which yields the smallest average (over item strata) estimated MSE. One may obtain an alternative estimator of r_{Ij} by subtracting $\sum_{i=1}^{I-1} r_{ij}$ from unity. However, according to the procedure, there is a positive probability that r_{Ij} estimate is negative.

4. NUMERICAL RESULTS

In this section, we evaluate various estimators of the cost weight and relative importance based on estimated mean squared error. We consider four rival estimators: the preliminary estimator, estimator proposed by Cohen and Sommers (1984), the estimator currently used by the BLS and the empirical Bayes estimator considered recently by Ghosh and Sohn (1990). The Cohen-Sommers estimator of the cost weight (before raking) is given by

$$\begin{aligned}
\hat{\theta}^{\mathrm{CS}}_{ij} &= \hat{\theta}^{\mathrm{CS}*}_{ij} \quad \text{if} \quad |\, \hat{\theta}^{\mathrm{CS}*}_{ij} - Y_{ij} \,| < \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \\
&= Y_{ij} + \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}^{\mathrm{CS}*}_{ij} \geq Y_{ij} + \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \\
&= Y_{ij} - \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \quad \text{if} \quad \hat{\theta}^{\mathrm{CS}*}_{ij} \leq Y_{ij} - \mathrm{c}\cdot\mathrm{sd}(Y_{ij})
\end{aligned}$$

where

$$\hat{\theta}^{\mathrm{CS}*}_{ij} = (1 - \hat{a}^{\mathrm{CS}}_{ij})Y_{ij} + \hat{a}^{CS}_{ij}Y_{i\cdot},$$

$$\hat{a}^{\mathrm{CS}}_{ij} = \min\left[1,\ (1 - N_j/N)\left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y_{ij})^2\right]\Big/\left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y^{(k)}_{i\cdot})^2\right]\right],$$

$$Y^{(k)}_{i\cdot} = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W^{(k)}_{jl}X_{ijl}\Big/\sum_{j=1}^{m}\sum_{l=1}^{n_j} W^{(k)}_{jl},$$

N_j = total number of consumer units in the population for the jth publication area,

$$N = \sum_{j=1}^{m} N_j,$$

$$\mathrm{sd}(Y_{ij}) = \sqrt{\left\{\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y_{ij})^2\right\}},$$

c = a safety factor determined by the BLS (see Table 2).

Table 2

Values of the Safety Factor *c* for the Major Areas

Major Area	NCNS	NCSR	NENS	NESR	SSNS	SSSR	WWNS	WWSR
	1	2	3	4	5	6	7	8
c	1.0	.5	1.0	.5	3.0	.25	1.0	.5

NCNS: North Central (Midwest) non-self-representing.
NCSR: North Central self-representing.
NENS: North East non-self-representing.
SSNS: South non-self-representing.
SSSR: South self-representing.
WWNS: West non-self-representing.
WWSR: West self-representing.

Their estimator for the relative importance is given by

$$\begin{aligned} \hat{r}_{ij}^{\mathrm{CS}} &= \hat{r}_{ij}^{\mathrm{CS}*} \quad \text{if} \quad |\,\hat{r}_{ij}^{\mathrm{CS}*} - R_{ij}\,| \le \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \\ &= R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{\mathrm{CS}*} \ge R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \\ &= R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \quad \text{if} \quad \hat{r}_{ij}^{\mathrm{CS}*} \le R_{ij} - \mathrm{c}\cdot\mathrm{sd}(R_{ij}), \end{aligned}$$

where

$$\hat{r}_{ij}^{\mathrm{CS}*} = (1 - \hat{d}_{ij}^{\mathrm{CS}})R_{ij} + \hat{d}_{ij}^{\mathrm{CS}} R_{i\cdot}^{\mathrm{CS}},$$

$$R_{i\cdot}^{\mathrm{CS}} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl} \Bigg/ \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl},$$

$$\begin{aligned} \hat{d}_{ij}^{\mathrm{CS}} &= \hat{d}_{ij}^{\mathrm{CS}*} \quad \text{if} \quad 0 < \hat{d}_{ij}^{\mathrm{CS}*} < 1, \\ &= 0 \quad \text{if} \quad \hat{d}_{ij}^{\mathrm{CS}*} \le 0, \\ &= 1 \quad \text{if} \quad \hat{d}_{ij}^{\mathrm{CS}*} \ge 1, \end{aligned}$$

$$\hat{d}_{ij}^{\mathrm{CS}*} = \frac{\dfrac{1}{20} \displaystyle\sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2 - \dfrac{1}{20} \displaystyle\sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})(R_{i\cdot}^{\mathrm{CS}(k)} - R_{i\cdot}^{\mathrm{CS}})}{\dfrac{1}{20} \displaystyle\sum_{k=1}^{20} (R_{ij}^{(k)} - R_{i\cdot}^{\mathrm{CS}(k)})^2},$$

$$R_{i\cdot}^{\mathrm{CS}(k)} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} \Bigg/ \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl},$$

$$\mathrm{sd}(R_{ij}) = \sqrt{\frac{1}{20}\sum_{k=1}^{20} (R_{ij}^{(k)} - R_{ij})^2}.$$

Since $\sum_{i=1}^{I} \hat{r}_{ij}^{\mathrm{CS}} \neq 1$, for our comparison purpose, we have divided $\hat{r}_{ij}^{\mathrm{CS}}$ by $\sum_{i=1}^{I} \hat{r}_{ij}^{CS}$.

The procedure currently used by the Bureau of Labor Statistics (see United States Department of Labor 1988) consists of a number of steps.

Step 1: Obtain an estimator of the cost weight as follows:

$$\hat{\theta}_{ij}^{\mathrm{CS}(1)} = \hat{r}_{ij}^{CS} \sum_{i=1}^{I} Y_{ij}.$$

Step 2: Final estimator of θ_{ij} is obtained from $\hat{\theta}_{ij}^{\mathrm{CS}(1)}$ using a ''raking'' procedure. The final estimator, denoted by $\hat{\theta}_{ij}^{\mathrm{BLS}}$, satisfies the following two conditions:

$$\sum_{i=1}^{I} \hat{\theta}_{ij}^{\mathrm{BLS}} = \sum_{i=1}^{I} Y_{ij},$$

$$\sum_{j=1}^{m} N_j \hat{\theta}_{ij}^{\mathrm{BLS}} = \sum_{j=1}^{m} N_j Y_{ij}.$$

Step 3: Finally an estimator for the relative importance is obtained as follows:

$$\hat{r}_{ij}^{\mathrm{BLS}} = \hat{\theta}_{ij}^{\mathrm{BLS}} \Big/ \sum_{i=1}^{I} \hat{\theta}_{ij}^{\mathrm{BLS}}.$$

In our numerical work, we have estimated N_j by $\sum_{l=1}^{n_j} W_{jl}$.

The MSE of an estimator e_{ij} of θ_{ij} is given by:

$$\begin{aligned} \mathrm{MSE} &= E(e_{ij} - \theta_{ij})^2 \\ &= E(e_{ij} - Y_{ij})^2 - V(Y_{ij}) + 2\,\mathrm{Cov}(e_{ij}, Y_{ij}), \end{aligned}$$

where it is assumed $E(Y_{ij} \mid \theta_{ij}) = \theta_{ij}$. The above formula is given in Cohen and Sommers (1984). As in the Ghosh and Sohn (1990) we estimate the three terms by the balanced repeated half samples available from the survey. For example,

$$E(e_{ij} - Y_{ij})^2 \doteq \frac{1}{20}\sum_{k=1}^{20} (e_{ij}^{(k)} - Y_{ij}^{(k)})^2,$$

$$V(Y_{ij}) \doteq \frac{1}{20}\sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2,$$

Table 3

Average Estimated MSE's for Different Estimators of θ_{ij}

Major Area	Average Estimated MSE of				
	Y_{ij}	$\hat{\theta}_{ij}^{\text{GS}}$	$\hat{\theta}_{ij}^{\text{CS}}$	$\hat{\theta}_{ij}^{\text{BLS}}$	$\hat{\theta}_{ij}$
NCNS	.020047	.011549 (22)	.009342 (53)	.014885 (25)	.009428 (52)
NCSR	.036620	.024783 (32)	.016017 (56)	.023627 (35)	.016155 (55)
NENS	.018162	.013299 (26)	.007327 (59)	.013046 (28)	.005504 (69)
NESR	.052883	.051100 (3)	.038911 (26)	.045610 (13)	.028958 (45)
SSNS	.021757	.013146 (39)	.009954 (54)	.014415 (33)	.006418 (70)
SSSR	.047500	.028984 (38)	.031743 (33)	.044238 (6)	.009270 (80)
WWNS	.052387	.029938 (42)	.017433 (66)	.030069 (42)	.010849 (79)
WWSR	.018223	.033529 (−83)	.009925 (45)	.014898 (18)	.005761 (68)

Note: The figures in the parenthesis represents percent improvement over the preliminary estimator, Y_{ij}.

$$\operatorname{Cov}(e_{ij}, Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - e_{ij})(Y_{ij}^{(k)} - Y_{ij}).$$

In the above $e_{ij}^{(k)}$ is the estimator e_{ij} based on the kth half sample ($k = 1, \ldots, 20$). For example,

$$\hat{\theta}_{ij}^{\text{CS}(k)} = (1 - \hat{a}_{ij}^{CS}) Y_{ij}^{(k)} + \hat{a}_{ij}^{\text{CS}} Y_{i\cdot}^{(k)},$$

$$\hat{\theta}_{ij}^{(k)} = (1 - \hat{a}_{ij}) Y_{ij}^{(k)} + \hat{a}_{ij} e_i' \hat{\mu}^{(k)}.$$

We obtain $\hat{\theta}_{ij}^{\text{BLS}(k)}$ by the multistep procedure used to obtain $\hat{\theta}_{ij}^{\text{BLS}}$ where we replace Y_{ij}, R_{ij}, $\hat{r}_{ij}^{\text{CS}}$ by $Y_{ij}^{(k)}$, $R_{ij}^{(k)}$ and $\hat{r}_{ij}^{\text{CS}(k)}$ respectively. Note that the above procedure does not take into account the variation due to the estimation of the coefficients (*i.e.*,a_{ij}'s) in the composite estimators. Cohen and Sommers (1984) recommended the use of half samples of half samples, or quarter samples to capture this additional variability. We could not use their procedure since our dataset did not contain these quarter samples.

The data we analyze arise out of 1982-83 Consumer Expenditure Survey (Diary survey). The expenditure class we consider is dairy products. There are in all six item strata in this class. They are (1) fresh whole milk, (2) other fresh milk and cream, (3) butter, (4) cheese, (5) ice cream and related products, and (6) other dairy products.

The MSE's of all the estimators considered are estimated for each publication area and item stratum. In Table 3 we report the average estimated MSE's of the estimators of θ_{ij}, the average being taken over all the item strata and all the publication areas within a major area. Notice that all the composite estimators except the one proposed by Ghosh and Sohn (1990) are better than the preliminary estimator for all the major areas in the average MSE sense. Both $\theta_{ij}^{\mathrm{CS}}$ and $\hat{\theta}_{ij}$ are better than $\hat{\theta}_{ij}^{\mathrm{BLS}}$. Our proposed estimator $\hat{\theta}_{ij}$ is better than $\hat{\theta}_{ij}^{\mathrm{CS}}$ in six out of eight major areas. In two major areas (NCNS and NCSR), $\hat{\theta}_{ij}^{\mathrm{CS}}$ is better than $\hat{\theta}_{ij}$, but the difference is very negligible.

In Tables 4 and 5, we try to demonstrate that the raking procedure may not be necessary. In Table 4, the parameter of interest is $\sum_{i=1}^{I}\theta_{ij}$, the true cost weight for the expenditure class. Here, due to the "raking" procedure, $\sum_{i=1}^{I}\hat{\theta}_{ij}^{\mathrm{BLS}} = \sum_{i=1}^{I} Y_{ij}$. We propose an alternative estimator as $\sum_{i=1}^{I}\hat{\theta}_{ij}$ and compare the average estimated MSE (over publication areas in a major area) with that of $\sum_{i=1}^{I} Y_{ij}$. In all the cases, we gain considerably.

Table 4

Average Estimated MSE's of Two Estimators of Average Consumer Expenditure for the Expenditure Class

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.12384	0.07969	36
NCSR	0.29819	0.13040	56
NENS	0.21658	0.07602	65
NESR	0.67486	0.20119	70
SSNS	0.21506	0.08303	61
SSSR	0.68415	0.06462	90
WWNS	0.35446	0.05175	85
WWSR	0.19292	0.05524	71

Table 5

Average Estimated MSE's of Two Estimators of Average Consumer Expenditure for the Major Area

Major Area	Preliminary Estimator	Proposed Estimator	Percent Improvement
NCNS	0.008181	0.0045468	44
NCSR	0.003672	0.0031047	15
NENS	0.006174	0.0029128	53
NESR	0.011680	0.0056922	51
SSNS	0.007501	0.0036401	51
SSSR	0.004434	0.0013751	69
WWNS	0.008203	0.0022560	72
WWSR	0.002786	0.0007882	72

In Table 5, the parameter of interest is the cost weight of an item stratum for the major area. The preliminary estimator (identical to the BLS estimator due to the raking procedure) is $(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl} Y_{ij})/(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. Our estimation procedure can also generate estimators at the major area level. We propose the estimator as $\hat{\theta}_{i\cdot} = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl}\,\hat{\theta}_{ij}/(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. The average estimated MSE's for these two estimators are reported in Table 5. Here also our estimator is superior to the preliminary (BLS) estimator.

The results of Table 4 and 5 suggest that the data consistency step followed by the BLS may not be necessary. Indeed, it may be possible to improve the traditional estimators at higher levels of aggregation also.

Table 6 provides the average estimated MSE's (over all the item strata and publication areas in a major area) of various estimators of relative importance. Notice that as in Table 3, all the estimators other than $\hat{r}_{ij}^{\mathrm{GS}}$ are better than the preliminary estimator $\hat{R}_{ij}$ for all the major areas. Our proposed estimator $\hat{r}_{ij}$ is the best among all the estimators considered.

Recently, Swanson (1992) has compared different methods of estimating cost weights for 12 of the approximately 70 expenditure classes in the CPI. His investigation shows that overall our proposed method is superior to all the rival methods.

Table 6

Average Estimated MSE's for Different Estimators of Relative Importance

Major Area	Average Estimated MSE of				
	R_{ij}	$\hat{r}_{ij}^{\mathrm{GS}}$	$\hat{r}_{ij}^{\mathrm{CS}}$	$\hat{r}_{ij}^{\mathrm{BLS}}$	$\hat{r}_{ij}$
NCNS	.0006342	.00046480 (27)	.00033143 (48)	.00042130 (34)	.00018592 (71)
NCSR	.0009125	.00071967 (21)	.00040226 (56)	.00044815 (51)	.00021309 (77)
NENS	.0003588	.00026894 (25)	.00014146 (61)	.0001620 (55)	.00011105 (69)
NESR	.0004264	.00072001 (−69)	.00028862 (32)	.00030555 (28)	.00016744 (61)
SSNS	.0005071	.00033736 (33)	.00019352 (62)	.00021385 (58)	.00011925 (76)
SSSR	.0006564	.00048569 (26)	.00053173 (19)	.00053603 (18)	.00030979 (53)
WWNS	.0013709	.00086849 (37)	.00051474 (62)	.00061901 (55)	.00028519 (79)
WWSR	.0003540	.00070770 (−100)	.00021384 (40)	.00023255 (34)	.00013750 (61)

Note: The figure given in the parenthesis represents percent improvement over R_{ij}.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation (NSF) under grant SES-9001399, "On-Site Research to Improve the Quality of Labor Statistics." This research was conducted at the U.S. Bureau of Labor Statistics while the authors were participants in the American Statistical Association/Bureau of Labor Statistics Research Program, which is supported by the Bureau of Labor Statistics and through the NSF grant. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of Labor Statistics. We wish to thank Sylvia Leaver, Stuart Scott, Malay Ghosh, Richard Valliant, Stephen Miller, So Young Sohn, Paul Hsen, Adriana Silberstein for many valuable discussions. We also thank two referees and associate editor for helpful comments on an earlier version of this paper.

REFERENCES

COHEN, M.P., and SOMMERS, J.P. (1984). Evaluation of methods of composite estimation of cost weights for the CPI. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 466-471.

GHOSH, M. and SOHN, S.Y. (1990). An Empirical Bayes Approach Towards Composite Estimation of Consumer Expenditure. Technical Report, U.S. Bureau of Labor Statistics.

MARKS , H. (1978). Composite estimation techniques used for the CPIR weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 311-315.

McCARTHY P.J. (1969). Pseudoreplication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

SWANSON, D. (1992). An evaluation of 4 cost weight composite estimation methods for the CPI. Memorandum for Janet Williams, Chief, CPI Survey Research Branch, Statistical Methods Division, U.S. Bureau of Labor Statistics.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2nd Edition). New York: J. Wiley & Sons.

UNITED STATES DEPARTMENT OF LABOR (1988). *Handbook of Methods*. Bureau of Labour Statistics. Washington DC: U.S. Government Printing Office.