Journal of Official Statistics, Vol. 16, No. 1, 2000, pp. 39–52

Optimal Weighting of Index Components: An Application to the Employment Cost Index

Michael K. Lettau and Mark A. Loewenstein[1]

This article considers the problem of how best to estimate an index whose components are themselves estimated with varying degrees of precision. We first characterize the weights that minimize the expected mean squared error in the index. These weights depend upon unobserved cell means and variances, but in practice one can substitute estimated cell means and variances for the unknown true parameters. We show that this is a more conservative reweighting scheme than one that minimizes a statistic that is an unbiased estimate of the mean squared error. We apply our procedure to the wage component of the Employment Cost Index (ECI) making use of the historical data that are used in the actual calculation of the ECI. While we apply our procedure to the ECI, our approach is equally valid for other indices.

Key words: Index number; composite estimation

1. Introduction

This article considers the problem of how best to estimate an index whose components are themselves estimated with varying degrees of precision. Let the index in period t be given by

$$I_t^* = \sum_{i=1}^{N} \omega_i \mu_i \tag{1}$$

where μ_{it} is the value of the i^{th} component and ω_i is the known, fixed weight given to the i^{th} component in the overall index. For example, I^* may be a price index in which case, μ_i is the proportionate change in price over the period for the i^{th} class of goods or services and ω_i is category i's share of the total budget. Naturally, if we knew the true values of the μ_i's, it would be straightforward to calculate the value of I^*. Unfortunately, we are only able to obtain an estimate of each μ_i. Depending on the nature of our sample, some of the μ_i's may be estimated much more precisely than others.

The following question immediately comes to mind: Might we want to over(under) weight those components that we are able to estimate most (least) precisely? More generally, what are the optimal weights? This article develops a procedure for solving precisely this problem. We then apply our procedure to the wage component of the Employment

[1] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 4130, Washington D.C. 20212, U.S.A. E-Mail: Lettau_M@bls.gov and Loewenstein_M@bls.gov

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. Department of Labor or the Bureau of Labor Statistics. We thank Larry Ernst, John Ruser, Alan Dorfman, several anonymous referees, and the editor for helpful comments.

Cost Index (ECI) making use of the historical data that are used in the actual calculation of the ECI. While we apply our procedure to the ECI, our approach is equally valid for other indices.

Our estimating technique can be thought of as an example of composite estimation. When we underweight a component in a price index because there is not enough information on good i to estimate its change in price very precisely, we are implicitly using information on the price changes of other goods to help estimate the change in good i's price.[2]

We should explicitly point out that sample sizes are treated as fixed in our analysis. Rather than adjusting the weights used to aggregate the various components of an index, it would generally seem preferable for the survey designer simply to choose sample sizes optimally to begin with. However, there are at least three reasons why this may not always be feasible. First, a survey may have competing uses. The survey design that minimizes mean squared error for one of the statistics of interest may not minimize the mean squared error for another. For example, the same establishment survey is used to produce both the Employment Cost Index, which is a Laspeyres index that measures changes in employers' cost of compensating workers after controlling for compositional changes in the distribution of employment, and the Employer Cost of Employee Compensation, which measures average compensation costs over time. Second, the stratification required by the ''optimal'' survey may be infeasible or very costly. For example, in the case of the ECI, it would be necessary to stratify by not only industry, but also by occupation. However, one does not know what particular occupations one will encounter at a given employer until that employer is interviewed. Third, one might obtain information on variances only by examining results that were obtained when a survey was fielded in the past and it may be too costly or even infeasible (in the case of a longitudinal survey) to change the survey design in the immediate future.

2. Optimal Weighting of Index Components

Because one does not directly observe the components μ_{it}, the index in Equation [1] is not observed, but rather must be estimated. Let the estimated value of the i^{th} component be given by

$$m_{it} = \mu_{it} + e_{it} \tag{2}$$

[2] Current Population Survey estimates are well-known examples of composite estimates. As noted by Bailar (1975), the composite estimates utilize the fact that in the CPS there is ''an overlap of about 75 percent of the persons in sample from one month to the next. If there is a fairly large positive correlation over time for persons in sample, (one) can make use of this to provide estimates with smaller variances.'' The composite estimate averages the simple level estimate for the current month with another estimate ''of level based on the composite estimate for the previous month to which has been added the estimated change from the preceding month to the current month'' and which ''is based on the part of the sample that is common to the two months.'' See Cantwell and Ernst (1992) for further discussion of composite estimation in the CPS, especially issues arising as a result of the 1994 procedural changes in the CPS. See Wolter (1979) for a good explanation of how composite estimators can be used in association with rotation schemes to reduce variances. Zieschang (1990) shows how composite estimation can lower variances of estimates obtained from the Consumer Expenditure Survey. Randolph and Zieschang (1988) derive composite estimators for local area price indices; when applied to estimate city level rental physical housing depreciation rates, these estimators result in large improvements in mean squared error.

where e_{it} is a random variable with mean 0 and variance σ_{it}^2 and where $E(e_{it}e_{jt}) = 0$ if $i \neq j$. Letting β_{it} denote the weight given to this component, the estimating index is given by

$$I_t = \sum_{i=1}^{N} \beta_{it}\mu_{it} \tag{3}$$

The expected mean squared error in the estimated index is defined by

$$MSE_t \equiv E\big((I_t - I_t^*)^2\big) \tag{4}$$

The optimal weights, β'_{it}, minimize (4). To find the optimal weights, substitute (1) and (3) into (4) and use (2), simplify to obtain

$$\begin{aligned} MSE_t &= E\left(\left(\sum_{i=1}^{N}\beta_{it}m_{it} - \sum_{i=1}^{N}\omega_i\mu_{it}\right)^2\right) \\ &= E\left(\left[\sum_{i=1}^{N}\beta_{it}e_{it} + \sum_{i=1}^{N}(\beta_{it}-\omega_i)\mu_{it}\right]^2\right) \\ &= \sum_{i=1}^{N}\beta_{it}^2\sigma_{it}^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_{it}-\omega_i)(\beta_{jt}-\omega_j)\mu_{it}\mu_{jt} \end{aligned} \tag{5}$$

Differentiating with respect to β_{it} yields the first-order conditions to the minimization problem:

$$\beta'_{it}\sigma_{it}^2 + \sum_{j=1}^{N}(\beta'_{jt} - \omega_j)\mu_{it}\mu_{jt} = \lambda \tag{6a}$$

$$\sum_{i=1}^{N}\beta'_{it} = 1 \tag{6b}$$

where λ is the Lagrange multiplier corresponding to the normalizing constraint that the β's sum to 0. To help interpret these conditions, let us rewrite (6a) as:

$$\beta'_{it}\sigma_{it}^2 + \mu_{it}(I'_t - I_t^*) = \lambda \tag{6a$'$}$$

where

$$I'_t = \sum_{j=1}^{N}\beta'_{jt}\mu_{jt}$$

The first term in (6a') captures the effect of an increase in β_{it} on variance, while the second term captures the effect of an increase in β_{it} on bias. If the weights are such that the new index exceeds (is less than) the true index, then the second term is positive and would cause β_i to be smaller (larger). The size of this depends on μ_{it}.

Equation (6) is a linear system of equations and can therefore be solved using Cramer's

rule. However, it is possible to express the solution in a form that is more useful:[3]

$$\beta'_{it} = \frac{1 + \sum_{j=1}^{N} (\mu_{it} - \mu_{jt})(\mu_t - \mu_{jt})/\sigma_{jt}^2}{\sum_{j=1}^{N} (\sigma_{it}^2/\sigma_{jt}^2) + \sum_{j=1}^{N} \sum_{k=1}^{N} \mu_{jt}(\mu_{jt} - \mu_{kt})(\sigma_{it}^2/\sigma_{kt}^2)/\sigma_{jt}^2} \tag{7}$$

where $\mu_t = \sum_j \mu_{jt}\omega_j$

Equation (7) indicates that the degree to which it is optimal to overweight and underweight the individual cells depends crucially on the differences in the cell means, $(\mu_i - \mu_j)$, and the variances σ_{it}^2.[4] In general, the smaller are the differences in the cell means relative to the variances, the greater is the over- and underweighting in response to suboptimal sample weights. In the extreme case where the cell means are all the same, Equation (7) reduces to

$$\beta'_{it} = \frac{\frac{1}{\sigma_{it}^2}}{\sum_{j=1}^{N} \left(\frac{1}{\sigma_{jt}^2}\right)}$$

so that each component's weight should be inversely proportional to the variance with which its value is estimated.[5] At the other extreme, it follows immediately from (6) that if the variances σ_{it}^2 are all 0, then $\beta'_{it} = \omega_i$, so that there is no over or underweighting.

To gain further insight into the optimal weighting scheme, consider the special case where there are only two cells. In this case, one can solve for β'_{it} to obtain

$$\beta'_{it} = \omega_i + \frac{\omega_j \sigma_{jt}^2 - \omega_i \sigma_{it}^2}{(\mu_{1t} - \mu_{2t})^2 + \sigma_{1t}^2 + \sigma_{2t}^2}. \tag{7'}$$

It is apparent from (7') that if $\sigma_{2t}^2/\sigma_{1t}^2 = \omega_1/\omega_2$, then $\beta'_{1t} = \omega_1$ and $\beta'_{2t} = \omega_2$. Thus, it is not optimal to overweight or underweight either component if $\sigma_{it}^2 = k/\omega_i$ for $i = 1$ and $i = 2$, that is, if the variance with which a component's value is estimated is inversely proportional to the component's proper weight in the index. Note, for example, that this condition will hold if component i's value is estimated as the average of n_i independent observations, provided that the variance associated with each observation is the same and n_i is proportional to ω_i. From (7a), one sees that if $\sigma_{2t}^2/\sigma_{1t}^2 > \omega_1/\omega_2$, then $\beta'_{1t} > \omega_1$ and $\beta'_{2t} < \omega_2$. Thus, if component

[3] Letting $x_{it} = \beta_{it} - \omega_i$ and $T = \Sigma_j x_{jt}\mu_{jt}$, Equation (6a) can be rewritten as

$$x_{it} + \omega_i + (\mu_{it}/\sigma_{it}^2)T - (\lambda/\sigma_{it}^2) = 0 \tag{a}$$

Multiplying (a) by μ_{it} and summing over i yields:

$$\left(1 + \sum_i \frac{\mu_{it}^2}{\sigma_{it}^2}\right) T + \sum_i \mu_{it}\omega_i - \lambda \sum_i \frac{\mu_{it}}{\sigma_{it}^2} = 0 \tag{b}$$

Summing (b) over i and using the fact that $x_{it} = \beta_{it} - \omega_i$ and $\Sigma_i \beta_{it} = 1$ yields

$$\sum_i \frac{\mu_{it}}{\sigma_{it}^2} T + \sum_i \omega_i - \lambda \sum_i \frac{1}{\sigma_{it}^2} \tag{c}$$

Solving (b) and (c) for T and λ and substituting into (a) yields Equation (7).

[4] While one might initially expect that the optimal weights β'_{it} should be expressible in terms of their distance from ω_i, note that the first term on the right hand side of (6a) involves β_{it} but not ω_i. This reflects the fact that other things being the same, variance can generally be reduced by shifting weight from higher weighted components to lower weighted components.

[5] Note that if the cell means are all the same, then the expected value of the index will be the same for all weighting schemes, but mean squared error will not.

1 is estimated relatively more precisely than component 2 after controlling for each component's importance in the index (i.e., if $\sigma_{1t}^2 = k_1/\omega_1$, $\sigma_{2t}^2 = k_2/\omega_2$, and $k_1 < k_2$), then it is optimal to overweight the first component and underweight the second. Note that this condition will obtain in our example above if $n_1/\omega_1 > n_2/\omega_2$.

While Equation (7) characterizes the weights that will minimize the mean squared error in the estimated index, this equation cannot be implemented in practice because the cell means, μ_i, are unknown (indeed, if the cell means were all known, one could calculate the value of the index directly from (1). One possible approach is simply to substitute the estimated cell means m_{it} for the unobservable true cell means μ_{it} in the first-order conditions (6a'), which gives us:

$$\beta_{it}\sigma_{it}^2 + m_{it}(I_t - \hat{I}_t) = \lambda \tag{6a''}$$

where $\hat{I}_t = \sum_{i=1}^{N} \omega_i m_{it}$

The weights solving (6a'') and (6b) are given by

$$\hat{\beta}_{it} = \frac{1 + \sum_{j=1}^{N}(m_{it} - m_{jt})(m_t - m_{jt})/\sigma_{jt}^2}{\sum_{j=1}^{N}(\sigma_{it}^2/\sigma_{jt}^2) + \sum_{j=1}^{N}\sum_{k=1}^{N} m_{jt}(m_{jt} - m_{kt})(\sigma_{it}^2/\sigma_{kt}^2)/\sigma_{jt}^2} \tag{8}$$

where $m_t = \sum_j m_{jt}\omega_j$

We will examine this solution more carefully subsequently. First, we address the problem of how to estimate mean squared error when the true cell means μ_i, are estimated with error.

3. Estimating Mean Squared Error

Recall from (5) that $MSE_t =$

$$\sum_{i=1}^{N} \beta_{it}^2\sigma_{it}^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_{it} - \omega_i)(\beta_{jt} - \omega_j)\mu_{it}\mu_{jt}.$$

Using the fact that $E(m_{it}^2) = \mu_{it}^2 + \sigma_{it}^2$ and $E(m_{it}m_{jt}) = \mu_{it}\mu_{jt}$ if $i \neq j$, we therefore have

$$MSE_t = E\left(\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_{it} - \omega_i)(\beta_{jt} - \omega_j)m_{it}m_{jt} + \sum_{i=1}^{N}\beta_{it}^2\sigma_{it}^2 - \sum_{i=1}^{N}(\beta_{it} - \omega_i)^2\sigma_{it}^2\right) \tag{9}$$

Thus, if

$\hat{\sigma}_{it}^2$

are unbiased estimates of the variances,

$$\hat{M}_t = \sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_{it} - \omega_i)(\beta_{jt} - \omega_j)m_{it}m_{jt} + \sum_{i=1}^{N}2(\beta_{it}\omega_i)\hat{\sigma}_{it}^2 - \sum_{i=1}^{N}\omega_i^2\hat{\sigma}_{it}^2 \tag{9'}$$

is an unbiased estimate of MSE_t.

Having derived a statistic to estimate mean squared error, let us return to our problem of choosing weights to minimize expected mean squared error. One possible approach is to choose weights to minimize the mean squared error statistic, $\hat{M}_t$. As shown below, the weights minimizing $\hat{M}_t$ depend on estimated variances and estimated cell means. Since our derivation of mean squared error required fixed cell weights that are independent of estimated cell means, the minimized value of $\hat{M}$ cannot be used as an estimate of mean squared error. There is no guarantee that the weights that minimize $\hat{M}_t$ will minimize expected mean squared error, but this certainly seems a reasonable approach to take. Furthermore, as discussed below, a comparison of the resulting first-order conditions with (6a'') yields insights into the weighting scheme (8) obtained by substituting estimating cell means for the true cell means in (7).

Differentiating (9′) with respect to β_{it} yields the first-order conditions that must be satisfied by the weights that minimize $\hat{M}_t$

$$\omega_i \sigma_{it}^2 + m_{it} \sum_{j=1}^{N} (\beta_{jt} - \omega_j) m_{jt} = \lambda \tag{10a}$$

$$\sum_{i=1}^{N} \beta_i = 1 \tag{10b}$$

This is a linear system of equations and is therefore theoretically solvable. However, if the number of cells is large, the standard linear solution is not empirically tractable. Nevertheless, it is possible to compare the weights that satisfy the first-order conditions (10a) with those that satisfy (6a''), obtained by substituting estimated cell means for the true cell means in (6a'). To help gain some intuition, let us first consider the case where there are only two cells.

When there are only two cells, the solution to (6a'') and (6b) can be rewritten as

$$\hat{\beta}_{it} = \omega_1 + \frac{\omega_j \sigma_{jt}^2 - \omega_i \sigma_{it}^2}{(\mu_{1t} - \mu_{2t})^2 + \sigma_{1t}^2 + \sigma_{2t}^2} \tag{8′}$$

In contrast, the solution to (10a) and (10b) when there are only two cells is given by

$$\hat{\hat{\beta}}_{it} = \omega_i + \frac{\omega_j \sigma_{jt}^2 - \omega_i \sigma_{it}^2}{(\mu_{1t} - \mu_{2t})^2} \tag{11}$$

Comparing (8') and (11), one sees that $|\hat{\beta}_{it} - \omega_i| < |\hat{\hat{\beta}}_{it} - \omega_i|$. That is, (8') constitutes a more conservative reweighting scheme than (11). A similar result should hold in the more general case. The system of equations (6a''), which is obtained by substituting the estimated cell means m_{it} for the unobservable true cell means μ_{it} in (6a'), does not account for the fact that we do not know the true index bias with certainty and thus overstates the effect of a change in β_1 on expected bias. To see this, note that the first-order conditions in (10a) can be rewritten as

$$\beta_{it} \sigma_{it}^2 + m_{it}(I_t - \hat{I}_t) + (\omega_i - \beta_{it}) \sigma_{it}^2 = \lambda \tag{10a′}$$

Equations (6a'') and (10a') are identical if $\beta_{it} = \omega_i$. Other things the same, a low value of

σ_{it}^2

will cause cell i to be overweighted, that is, cause β_{it} to exceed ω_i. But for $\beta_{it} > \omega_i$, the left-hand side of (6a'') will exceed the right-hand side of (10a'): when we ignore the fact that the bias is estimated with uncertainty, we will overstate the cost of overweighting beta. Similarly, for $\beta_{it} < \omega_i$, the left-hand side of (6a'') will be less than the right-hand side of (10a'): when we ignore the fact that the bias is estimated with uncertainty, we will understate the gain to underweighting beta. Thus, the weights satisfying (6a'') constitute a more conservative reweighting scheme than those satisfying (10a').

4. An Application Using the Employment Cost Index

We now illustrate our proposed procedure with actual data used to calculate the Employment Cost Index. The Employment Cost Index or ECI measures changes in employers' cost of compensating workers, controlling for changes in the industrial-occupational composition of jobs. The index is calculated using a two-step aggregation procedure. The first step aggregates microdata from individual quotes to estimate compensation for approximately 650 categories of labor, where the categories of labor are defined by pseudo industry (PSIC) and major occupation group (MOG). The PSICs correspond approximately to 2-digit SIC industries. The second step aggregates these PSIC/MOG cells to form the index.

Regarding the first step of the aggregation, if a cell does not have a sufficient number of nonimputed quotes, it is collapsed to the cluster/MOG level. A cluster is a group of two to five PSICs. For example, if a cell contains only one quote, but the cluster/MOG cell contains four additional quotes, compensation is averaged among the five quotes, and this average is then applied to the PSIC/MOG cell with only one quote. If the entire cluster/MOG cell does not contain a sufficient number of quotes, the cell is collapsed to a higher level of aggregation, up to the major industry group (MIG)/MOG level, which is the highest level of aggregation used. Note that this collapse procedure implicitly defines a weighting scheme for the ECI.[6] While the current procedure is not unreasonable, it certainly has an element of arbitrariness.

One would like to compare alternative aggregation schemes. Toward this end, we assume that the natural log of the ratio of the average hourly earnings in quarter t to quarter $t-1$ for the j^{th} job in the i^{th} PSIC/MOG cell is given by

$$\ln(w_{ijt}/w_{ijt-1}) = \mu_{it} + \epsilon_{ijt} \tag{12}$$

where ϵ_{ijt} is a random variable with mean 0 and variance σ_{it}^2. Letting ω_i denote cell i's share of the total wage bill in the base year, where each cell is defined by a unique industry-occupation combination, the true ECI is given by Equation (1).[7] Letting β_i denote the actual weight given to the i^{th} cell in the estimated ECI and letting m_{it} denote

[6] For example, suppose for simplicity that the index has only two components. Suppose the first cell only has one quote and the second cell has four quotes and suppose that all quotes have the same sample weight. If average compensation in the first cell is estimated as the average of all five quotes - that is, if the first cell is collapsed into the second - then in effect cell 1 is only accorded one fifth of its true weight, so that $\beta_1 = (1/5)\omega_1$ and $\beta_2 = \omega_2 + (4/5)\omega_1$. Thus, the first (second) component is under (over) weighted in the estimated index.

[7] In practice, the ECI is calculated as an arithmetic mean of average wage changes. We use the geometric form because it is more convenient to work with. For a detailed description of the ECI, see BLS Handbook of Methods (1997), U.S. Bureau of Labor Statistics.

the sample mean of $\ln(w_{it}/w_{it-1})$ for those jobs in the i^{th} cell, the estimated index is given by (3). Letting n_i denote the number of sample observations in cell i, m_{it} has mean μ_{it} and variance σ_{it}^2/n_{it}.

It is straightforward to estimate σ_{it}^2. Specifically, Equation (10) can be estimated by regressing the change in the log wage rate on PSIC/MOG dummy variables. One can then estimate the within-cell variances σ_{it}^2 from the residuals to this regression. Denoting this estimated residual by $\hat{\sigma}_{it}^2$, our estimate of the mean squared error is

$$\hat{M}_t = \sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_{it}-\omega_i)(\beta_{jt}-\omega_j)m_{it}m_{jt} - \sum_{i=1}^{N}2(\beta_{it}-\omega_i)\frac{\hat{\sigma}_{it}^2}{n_{it}} + \sum_{i=1}^{N}\omega_i^2\frac{\hat{\sigma}_{it}^2}{n_{it}}$$

Using data from 1990–1994, we have calculated the expected mean squared error estimate, $\hat{M}_t$, for the procedure that is currently being used. Averaging over all periods, we find that the average mean squared error associated with the current weighting scheme is $0.574 \cdot 10^{-6}$. Table 1 also presents the average mean squared error associated with three other possible weighting schemes: a scheme that never collapses and schemes that always aggregate up to the cluster/MOG and MIG/MOG level, respectively.[8] The more aggregated definitions for the categories of labor in the first step of the ECI aggregation largely eliminate the need for a formal collapse procedure, thus simplifying the calculation of the ECI.

Of the four schemes, the current procedure produces the lowest estimated mean squared error at $0.574 \cdot 10^{-6}$, which is slightly lower than the estimated, mean squared error under no collapsing. The mean squared errors are slightly larger for the more aggregated schemes. To put these estimates in context, we also report the average value of the mean squared error normalized by the ECI estimate,

$$\frac{\sum_t \frac{\sqrt{Z_t}}{ECI_t}}{T}$$

This estimated statistic is very similar for all four of the aggregation schemes. For additional insight into this result, we have estimated the expected squared difference between a PSIC/MOG cell mean and its MIG/MOG cell mean. The estimate basically equals the average of the squared difference between a PSIC/MOG cell mean and the mean outside the cell but within its MIG/MOG, with an adjustment for the fact that the means are estimated rather than known with certainty.[9] We estimate the parameters separately for all quarters from March 1990 through December 1994, with the restriction that each PSIC/MOG cell must have at least two nonimputed quotes and each higher-level cell must have at least two nonimputed quotes outside each of its PSIC/MOG cells. The resulting parameter estimates appear in Table 2. The last column of

[8] We use wage and salary data only, and we restrict the sample to the private sector. In actual practice, collapsing cannot be avoided when a cell is empty. However, since our present concern is with cells that have a small sample size but are not empty, we have dropped empty cells and reallocated their weights to all nonempty cells in proportion to their size. This has very little effect on our estimates of mean squared error. In obtaining the mean squared errors in Table 1, we have assumed that the within-cell variance σ_{it}^2 is the same for all cells. However, we have also obtained separate estimates for each of the nine major occupational groups, in the process allowing within-cell variances to differ across the major occupational groups. When one takes employment weighted averages across the occupational groups, one obtains mean squared error estimates that are virtually identical to those in Table 1.

[9] A more detailed description of this estimation can be found in the Appendix.

Table 1. Mean squared error estimates for alternative aggregation schemes

	ECI	Mean squared error	Normalized root Mean squared error
No collapsing	0.007880	$0.588 \cdot 10^{-6}$	0.1016
Current procedure	0.007909	$0.574 \cdot 10^{-6}$	0.0998
Cluster/MOG	0.007893	$0.652 \cdot 10^{-6}$	0.1052
MIG/MOG	0.007901	$0.668 \cdot 10^{-6}$	0.1040

Table 2 refers to the test statistic for the joint restriction that the means for all PSIC/MOG cells equal the means for their corresponding MIG/MOG cells. The marginal significance shows the lowest level of significance at which this null hypothesis is rejected. It is clear from the table that the expected squared difference between a PSIC/MOG cell mean and its MIG/MOG cell mean, while not necessarily zero, is certainly quite small.

The estimating procedures discussed above reallocate weights within MIG/MOG cells but not outside of them. An alternative, more aggressive approach is to use the weighting scheme described by (8). In this approach, weights are reallocated across MIG/MOG cells as well as within them; indeed, the decision to reallocate a weight from one PSIC/MOG cell to another is completely independent of whether or not the two cells are in the same MIG/MOG. When one uses these weights, the value of $\hat{M}_t$ turns out to be $0.039 \cdot 10^{-6}$, which is markedly lower than the mean squared error associated with the current procedure or any of the other weighting schemes listed in Table 1. However, this is not a meaningful comparison. Our mean squared error derivation requires fixed cell weights that are independent of estimated cell means. The weighting schemes listed in Table 1 all satisfy this requirement. The weights in (8) clearly do not.

Calculating the mean squared error for this scheme is very difficult because one must take into account variances in the weights caused by variances in the mean squared estimates. As an alternative, we perform a Monte Carlo simulation. Using the parameter estimates in Table 2, we simulate a set of sample quotes for each of the twenty quarters.[10] We next determine the resultant weights for the current procedure and the weights described by (8). Since we know the underlying parameters generating the simulated sample quotes, we can then calculate both the $\hat{M}_t$ statistics and the true expected mean squared errors for the two aggregation schemes using Equation (5).

We have performed 60 simulations, where each simulation itself involves simulating sample quotes for each of the twenty quarters from 1990–1994. As expected, the average value of $\hat{M}_t$ for the current procedure, $0.582 \cdot 10^{-6}$, is very close to the average value of the true mean squared error of $0.583 \cdot 10^{-6}$, verifying that $\hat{M}_t$ is a consistent estimate of the expected mean squared error. As we also suspected, $\hat{M}_t$ is an underestimate of true mean squared error for our alternative procedure that chooses weights according to (8). The average minimized value of $\hat{M}_t$ is again quite small at $0.011 \cdot 10^{-6}$, while the true mean squared error for this procedure is $0.390 \cdot 10^{-6}$. This mean squared error is only

[10] Besides the parameter estimates in Table 2, the simulation also utilizes estimates of the variance of wage growth across MIG/MOGs. Our simulated set of sample quotes therefore has similar means, variances, and sample sizes to those that are observed in the actual data.

Table 2. Parameter estimates for composite estimation. PSIC/MOG cells versus MIG/MOG cells

Quarter	No. of PSIC/ MOG Cells	No. of MIG/ MOG Cells	Within PSIC/ MOG Variance	Within MIG/ MOG Variance	Bias Squared	Adjusted Bias Squared	Marginal Significance of Test Statistic
9003	500	68	0.00451	0.00454	0.00047	0.00017	2.1%
9006	498	66	0.00396	0.00403	0.00041	0.00015	0.0%
9009	498	68	0.00370	0.00370	0.00027	0.00002	0.8%
9012	496	68	0.00321	0.00323	0.00026	0.00004	0.0%
9103	488	67	0.00348	0.00346	0.00020	−0.00002	70.2%
9106	492	68	0.00629	0.00622	0.00028	−0.00014	99.8%
9109	486	67	0.00356	0.00362	0.00028	0.00006	0.0%
9112	480	68	0.00299	0.00303	0.00024	0.00005	14.7%
9203	475	68	0.00359	0.00367	0.00029	0.00007	0.1%
9206	468	67	0.00292	0.00296	0.00024	0.00006	1.9%
9209	461	66	0.00336	0.00350	0.00046	0.00026	0.0%
9212	454	66	0.00317	0.00316	0.00019	−0.00001	99.6%
9303	456	66	0.00300	0.00299	0.00018	−0.00001	84.0%
9306	449	63	0.00252	0.00252	0.00015	−0.00001	27.1%
9309	441	61	0.00283	0.00281	0.00013	−0.00004	57.9%
9312	446	61	0.00257	0.00253	0.00011	−0.00005	99.8%
9403	440	57	0.00362	0.00362	0.00023	0.00000	2.1%
9406	422	58	0.00288	0.00296	0.00035	0.00016	0.0%
9409	409	59	0.00309	0.00307	0.00020	−0.00004	6.3%
9412	406	59	0.00314	0.00317	0.00029	0.00007	0.0%
Average			0.00342	0.00244	0.00026	0.00004	

about two-thirds as large as the expected mean squared error associated with the current procedure. The weighting scheme in (8) apparently results in an index that has a considerably lower mean squared error than the current procedure.

5. Conclusions

This article has examined the problem of how best to estimate an index whose components are themselves estimated with varying degrees of precision. Given sufficient information on the underlying parameters, we have characterized the component weights that minimize the mean squared error in the estimated index. Although the true mean squared error cannot be observed in practice, we implement our procedure by replacing unobserved cell means and variances by their estimated parameter values. This is a more conservative reweighting scheme than one that minimizes a statistic that is an unbiased estimate of the mean squared error.

We have illustrated the application of our procedure using actual ECI historical data. Interestingly, our small estimates for the bias between the PSIC/MOG cells and the higher-level cells suggest that the ECI collapse hierarchy is reasonable. Our results further indicate that a procedure that aggregates up to the MIG/MOG level performs nearly as well as the current ECI procedure. An advantage to the MIG/MOG aggregation scheme is that it virtually eliminates the need for any collapse procedure and thus simplifies the

calculation of the ECI. The current procedure and a procedure that aggregates up to the MIG/MOG reallocate weights within MIG/MOG cells, but not outside of them. A procedure that chooses weights to approximately minimize mean squared error is a more aggressive approach, since weights are reallocated across MIG/MOG cells as well as within them. Our results indicate that this procedure yields an index that has a substantially lower expected mean squared error.

While we have applied our procedure to the ECI, our approach is equally valid for other indices. Note also that we have only considered an ECI whose categories of labor are defined by PSIC/MOG cells, and we have only considered collapse schemes that aggregate across industry clusters. A more exhaustive evaluation would start with no assumption as to how individual quotes should be combined into categories of labor. Our procedure for estimating within cell variance and the expected bias when one uses a higher level aggregate to estimate a lower level cell mean and our derivation of the expected mean squared error associated with an arbitrary aggregation scheme provides the necessary theoretical framework for such an analysis.

Finally, we might note that our approach to calculating a price index is similar in spirit to what has become known as the stochastic approach to index number theory, which, as described by Selvanathan and Prasada Rao (1994), '' considers the index number problem as a signal extraction problem from the messages concerning price changes for different commodities.''[11] While the economic approach to index numbers assigns a weight to a price relative on the basis of its economic importance, the stochastic approach assigns the weight to a price relative on the basis of the strength of its signal. Diewert (1995) has criticized the stochastic approach on the grounds that (1) the variance assumptions that are made are not consistent with the observed behavior of prices and (2) ''if price relatives are different, then an appropriate definition of average price cannot be determined independently of the economic importance of the corresponding goods'' (p. 21). Neither of these criticisms applies to the approach that we have presented. Our variances are driven by the data on hand. Furthermore, our idealized weights in the absence of sampling error are presumably those suggested by economic theory. Sampling considerations cause our estimating weights to deviate from these economic weights, with the result that our approach essentially combines elements of both the economic and the stochastic approaches.

Diewert (1995) concludes that ''in the present context where all prices and quantities are known without sampling error, signal extraction approaches to index number theory should be approached with some degree of caution,'' but he goes on to note that ''there is a huge role for statistical approaches to index numbers when we change our terms of reference and assume that the given price and quantity data are only samples.'' Of course, this is precisely the motivating factor behind the approach we have taken in this article.

[11] The stochastic approach to index numbers dates back to Jevons (1865) and Edgeworth (1888). Modern contributions to the literature include Balk (1980), Clements and Izan (1987), Bryan and Cecchetti (1993), and Selvanathan and Prasada Rao (1994). Diewert (1995) provides a critical review of the literature.

Appendix

Equation (A1) puts the data from individual quotes into the framework of a regression model.

$$y = X\gamma + \epsilon \tag{A1}$$

where: $y = n \times 1$ vector of $\ln(w_{ijt}/w_{ijt-1})$
$X = n \times k$ matrix of dummy variables for the PSIC/MOG cells
$\gamma = k \times 1$ vector of μ_{jt}
$\epsilon = n \times 1$ vector of residuals

We estimate the parameters of Equation (A1) separately for each quarter t, so we drop the t subscript in the matrix definitions and in subsequent equations. Sample weights are normalized to sum to one. To simplify the ensuing discussion, we assume homoskedasticity throughout; experimentation with the date indicates that our results do not appear to be very sensitive to this assumption. The variable n refers to the number of quotes, and k refers to the number of PSIC/MOG cells. Let μ_j^c denote the expected wage growth of a quote that is outside of cell j but in cell j's MIG/MOG.

Define a matrix $M = I - X\Delta(X'\Omega X)^{-1}X'\Omega$, where the matrix I is an $n \times n$ identity matrix and Ω is an $n \times n$ diagonal matrix such that the element in the jth row and jth column is quote j's sample weight. The matrix Δ is a $k \times k$ matrix. The j,k^{th} element of Δ is zero if $j = k$ or if cell k is outside of cell j's MIG/MOG. If $j \neq k$ and if cell k is in cell j's MIG/MOG, then the j,k^{th} element of Δ equals cell k's sample weight divided by the total sample weight of all cells other than cell j that are in j's MIG/MOG.

Premultiplying Equation (A1) by M yields

$$My = MX\gamma + M\epsilon \tag{A2}$$

Note that My is an $n \times 1$ vector, the jth entry of which is simply the amount by which quote j's wage growth exceeds the average wage growth of all quotes that are in quote j's MIG/MOG but are not in quote j's cell. Note also that since $MX = X(I - \Delta)$, Equation (A2) can be rewritten as

$$My = X\beta + M\epsilon \tag{A3}$$

where β is a $k \times 1$ vector whose jth entry is $\mu_j - \mu_j^c$.

Let the $k \times 1$ vector b be the weighted least squares estimate of β and let e be the $nx1$ vector of residuals. The regression model provides natural estimates for the variance and bias squared:

$$E[(e'\Omega e)] = \sigma^2\{1 - \text{trace}[X'\Omega\Omega X(X'\Omega X)^{-1}]\} \tag{A4}$$

$$\begin{aligned} E[Xb)'\Omega(Xb)] = {} & \Sigma_j \pi_j(\mu_j - \mu_j^c)^2 + \sigma^2\{\text{trace}[X'\Omega\Omega X(X'\Omega X)^{-1}] \\ & + \text{trace}[\Delta' X'\Omega X\Delta\ (X'\Omega X)^{-1}X'\Omega\Omega X(X'\Omega X)^{-1}]\} \end{aligned} \tag{A5}$$

where the variable π_j equals the sum of the sample weights among quotes from cell j. The estimate for the within-cell variance corresponds to the usual least-squares estimate of the residual variance. (Note that if the sample weights all equal $1/n$, the trace of

$[X'\Omega\Omega X(X'\Omega X)^{-1}]$ equals k/n, so that we have the usual $n-k$ correction for the degrees of freedom in an ordinary least-squares regression.) The estimate for the bias squared equals the weighted average of the squared difference between the sample mean for the cell and the sample mean inside the cluster but outside the cell, with an adjustment for the fact that means are estimated rather than known with certainty.

We estimate the variance within a MIG/MOG but outside a PSIC/MOG cell (σ_j^{2c}) using the residual variance from the regression Equation (A6), which is the same as Equation (A1) except that the PSIC/MOG dummy variables are replaced by MIG/MOG dummy variables.

$$y = X^{cl}\gamma^{cl} + \eta \tag{A6}$$

where: $X^{cl} = n \times k^{cl}$ matrix of dummy variables for the cluster MOG cells.
$\gamma^{cl} = k^{cl} \times 1$ vector of parameters
$\eta = n \times 1$ vector of residuals
k^{cl} = the number of cluster/MOG cells

Finally, we use Equation (A7) to test the joint restriction that the means for all PSIC/MOG cells equal the means for their corresponding MIG/MOG cells.

$$y = X^{cl}\gamma^{cl} + X^{c}\gamma^{c} + \epsilon \tag{A7}$$

where: $X^c = n \times (k - k^{cl})$ matrix of dummy variables for the PSIC/MOG cells, after dropping one variable from each MIG/MOG cell.
$\gamma^c = (k - k^{cl}) \times 1$ vector of parameters.

Equation (A7) is equivalent to Equation (A1), except the parameters are redefined so that the joint restriction is equivalent to $\gamma^c = 0$.

6. References

Bailar, B. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. Journal of the American Statistical Association, 70, 23–30.

Balk, B.M. (1990). A Method for Constructing Price Indices for Seasonal Commodities. Journal of the Royal Statistical Society, A, 143, 68–75.

Bryan, M.F. and Cecchetti S.G. (1993). The Consumer Price Index as a Measure of Inflation. Economic Review, Federal Reserve Bank of Cleveland, 29, 15–24.

Cantwell, P. and Ernst, L. (1992). New Developments in Composite Estimation for the Current Population Survey. Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys, 121–130.

Clements, K.W. and Izan, H.Y. (1987). The Measurement of Inflation: A Stochastic Approach. Journal of Business and Economic Statistics, 5, 339–350.

Diewert, W.E. (1995). On the Stochastic Approach to Index Numbers. Discussion Paper #DP95-31, University of British Columbia.

Edgeworth, F.Y. (1888). Some New Methods of Measuring Variation in Generally Prices. Journal of the Royal Statistical Society, 51, 346–368.

Jevons, W.S. (1865) The Variations of Prices and the Value of the Currency since 1782.

Journal of the Royal Statistical Society of London, 28, 294–320; reprinted in Investigations in Currency and Finance. London: MacMillan and Co., 1884, 119–150.

Randolph, W. and Zieschang, K.D. (1988). Aggregation Consistent Restriction Based Improvement of Local Area Estimators. Working Paper #182, U.S. Bureau of Labor Statistics.

Selvanathan, E.A. and Prasada Rao, D.S. (1994). Index Numbers: A Stochastic Approach. Ann Arbor: The University of Michigan Press.

U.S. Bureau of Labor Statistics. (1997). BLS Handbook of Methods. U.S. Department of Labor, Bulletin 2490.

Wolter, K.M. (1979). Composite Estimation in Finite Populations. Journal of the American Statistical Association, 74, 604–613.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. Journal of the American Statistical Association, 986–1001.

Received February 1998
Revised April 1999