Use of Auxiliary Information to Evaluate a Synthetic Estimator in the U.S. Current Employment Statistics Program

J. Gershunskaya , J.L. Eltinge, and L. Huff, BLS
J. Gershunskaya, Statistical Methods Division, OEUS, gershunskaya_j@bls.gov
PSB 4985, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

Key Words: Covered Employment and Wages (ES-202) Program; Mean squared prediction error; Small area estimation; Small domain estimation; Nonresponse; Reporting error

Abstract:

The Bureau of Labor Statistics has considerable interest in the estimation of total monthly employment for small domains defined by the intersection of metropolitan statistical areas (MSA) and major industrial divisions (MID), based on data from the Current Employment Statistics Survey (CES). One of several possible elementary estimators is a synthetic estimator based on state-level changes in employment within a major industrial division. It is important to evaluate empirically the magnitude of the bias of this estimator, relative to the magnitude of the standard error of this estimator, and relative to the magnitudes of the biases and standard errors of other candidate elementary small domain estimators. This paper studies the extent to which this type of evaluation may be enhanced through the use of auxiliary data from the Covered Employment and Wages (ES-202) Program, a nominal census of employment that provides data several months after production of CES estimates. Principal attention is devoted to evaluation of components of mean squared error attributable, respectively, to: (1) lack of fit in the implicit synthetic model; (2) sampling error in the CES data; and (3) nonsampling error in the CES data.

1. Introduction

<u>ES-202 and CES Programs.</u> The Covered Employment and Wages program, also known as ES-202, is a cooperative program between the Bureau of Labor Statistics and the State Employment Security Agencies (SESA). Each covered employer is assigned an Unemployment Insurance (UI) account by the SESA. Each SESA collects monthly employment and total quarterly wages each quarter for all of the UI accounts within their respective state, and the files from this effort are made available to BLS approximately 6 to 9 months after the end of the reporting quarter.

The Current Employment Statistics (CES) program is a nationwide survey of business establishments that collects total payroll employment, monthly hours paid, and total payroll. The data is collected monthly from a sample of employers. The main advantage of the CES survey over the ES-202 is the timeliness of the data.

Since the ES-202 data consist of virtually all employers in the U.S., the data are considered "truth" for employment. The ES-202 data can be aggregated for any level of area (county and above) and industrial detail. The CES employment estimates are benchmarked or adjusted to the ES-202 levels once each year. In addition, the UI records collected under the ES-202 program provide a sampling frame for the CES survey as well as all other BLS establishment based surveys.

Availability of the true values from the ES-202 on a lagged basis provides opportunity for empirical evaluation of the CES estimators.

<u>Sample selection for the CES.</u> The CES sample is selected once a year as a stratified probability sample of UI accounts. The stratification is based on state, 11 major industrial divisions (MID), and 8 employment size classes. Allocation minimizes variance of month-to-month change at a state level at a fixed cost per state. See Bureau of Labor Statistics (1997), Butani et al. (1997), Werking (1997) for more information on the sample selection process.

<u>Variance Estimation.</u> Balanced repeated replication (BRR) methodology is used to produce variances. In defining strata for variance estimation, the three largest employment size classes are collapsed into one size class.

<u>Small domain estimation.</u> States have a need to produce employment estimates at very detailed levels of industry and geography. The design of the CES sample does not ensure adequate sample size for smaller domains to make reliable estimates directly from the sample. Small domain estimation methods are currently being

explored in connection with producing estimates from the CES sample for small intersections of MID and metropolitan statistical area (MSA).

The Estimators. The CES uses a weighted link relative (WLR) estimator. For a domain d at month t an estimate is defined as

$$\hat{Y}^{CES}_{d,t} = \hat{Y}^{CES}_{d,t-1}\hat{R}^{CES}_{t},$$

where growth rate

$$\hat{R}^{CES}_{t} = \frac{\sum_{i\in M_t} w_i y^{CES}_{i,t}}{\sum_{i\in M_t} w_i y^{CES}_{i,t-1}}$$

$\hat{Y}^{CES}_{d,t-1}$ = estimate for a domain d at month $t-1$; M_t = set of units reporting nonzero employment in both months t and $t-1$; w_i = selection weight of a sample unit i; $y^{CES}_{i,t}$, $y^{CES}_{i,t-1}$ = employment reported to the CES by unit i in the respective months.

Once a year, at month $t=0$, estimates are aligned with a base level obtained from the ES-202 for the domain d.

M_t for a direct estimator consists of units that belong to the domain d. M_t of a synthetic estimator contains larger set of units: when d is defined as an intersection of MID and MSA, M_t is a set of units from a statewide MID. Note that

$$\hat{Y}^{CES,Synthetic}_{d=MID*MSA,t} = \hat{Y}^{CES,Direct}_{d=MID,t} * F^{ES-202}_{MID,MSA,t=0}, \quad (1)$$

where $F^{ES-202}_{MID,MSA,t=0} = Y^{ES-202}_{MID*MSA,t=0} \big/ Y^{ES-202}_{MID,t=0}$

Sources of Error. There are several potential sources of error for the CES employment estimates.

Birth and death of establishments cause the frame to be imperfect. Small sample sizes generally lead to large *sampling errors*. There may be *nonresponse bias* due to responding units generally behaving differently than the nonresponding units. Low response rates may cause higher levels of nonresponse bias. Another important source of bias may arise because the employment data reported to the CES and the ES-202 are different. This is viewed as *reporting or measurement error* in the CES reported employment values because, by definition, the ES-202 employment values are considered truth and are used as the universe employment total.

For the synthetic estimator, another bias problem may arise because of the difference in the MID employment growth rate in the MSA of interest and the MID employment growth rate statewide.

The knowledge of patterns in these sources of error would possibly give directions in improving data collection strategy or make appropriate bias adjustments to the estimates.

This paper presents results from an empirical study of possible patterns and magnitude of error in the CES estimates. The study uses ES-202 data that become available several months after production of the CES estimates.

2. Error Decomposition Approach

Let $\hat{Y}^{Sample}_{d,t}$ be a full sample estimator (assuming a 100% response rate) computed using the ES-202 data. Sample or frame problems can be assessed by comparison of $\hat{Y}^{Sample}_{d,t}$ to the true ES-202 level $Y^{ES-202}_{d,t}$:

$$\hat{Y}^{Sample}_{d,t} = Y^{ES-202}_{d,t} + Sampling Error$$

Let $\hat{Y}^{Respondents}_{d,t}$ denote an estimator computed from the ES-202 for units that respond to the CES. The comparison of $\hat{Y}^{Sample}_{d,t}$ to $\hat{Y}^{Respondents}_{d,t}$ will tell us about the nonresponse error:

$$\hat{Y}^{Respondents}_{d,t} = \hat{Y}^{Sample}_{d,t} + Nonresponse\ Error$$

In addition, the CES estimate is a subject to reporting error:

$$\hat{Y}^{CES}_{d,t} = \hat{Y}^{Respondents}_{d,t} + Reporting Error$$

Overall error can be assessed from the comparison of $\hat{Y}^{CES}_{d,t}$ to $Y^{ES-202}_{d,t}$:

$$\hat{Y}^{CES}_{d,t} = Y^{ES\text{-}202}_{d,t} + Overall Error$$

Figure 1 displays a time plot of the ES-202 values and ES-202 values adjusted for establishment birth and death, as well as point estimates and confidence bounds computed from $\hat{Y}^{Sample}_{d,t}$, $\hat{Y}^{Respondents}_{d,t}$ and $\hat{Y}^{CES}_{d,t}$, respectively.

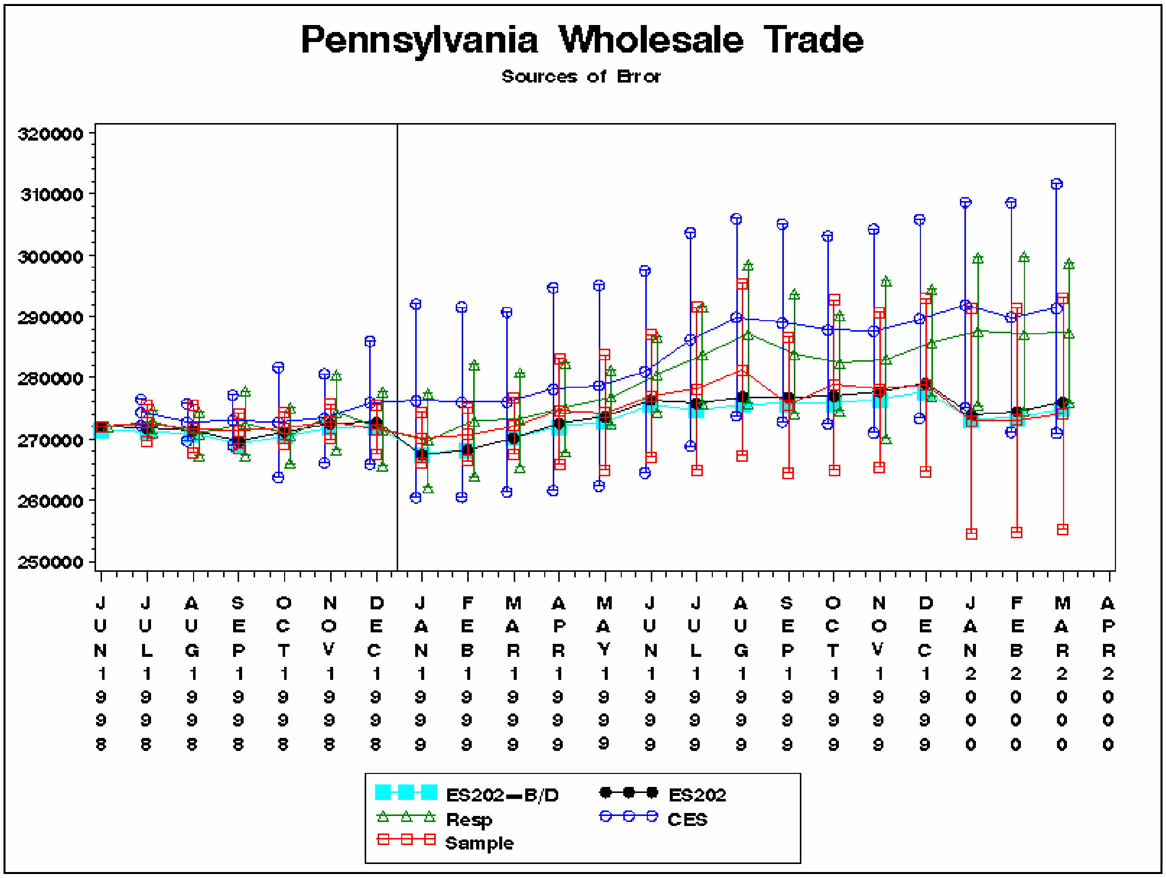

Figure 1. Comparison of the ES-202 and Birth-Death-Adjusted ES-202 (ES202-BD) with $\hat{Y}^{Sample}_{d,t}$ (Sample), $\hat{Y}^{Respondents}_{d,t}$ (Resp), and $\hat{Y}^{CES}_{d,t}$ (CES) for Pennsylvania State Wholesale Trade direct estimate

The CES sample used in the example was selected using first quarter 1998 ES-202 data. The estimates are aligned to the ES-202 level for Pennsylvania wholesale trade in June 1998. The vertical line shown on the graph indicates the beginning of the period when CES estimates from this sample would be published. In this particular example, the error due to nonresponse seems to be most significant.

Noting the relationship (1), we, at first, assess error components of the direct estimator of a statewide MID $\hat{Y}^{CES}_{d=MID,t}$. Define relative sample error

$$RelSE_{d,t} = 100\% \frac{\hat{Y}^{Sample}_{d,t} - Y^{ES-202}_{d,t}}{Y^{ES-202}_{d,t}},$$

relative nonresponse error

$$RelNR_{d,t} = 100\% \frac{\hat{Y}^{Respondents}_{d,t} - \hat{Y}^{Sample}_{d,t}}{\hat{Y}^{Sample}_{d,t}},$$

relative reporting error

$$RelRE_{d,t} = 100\% \frac{\hat{Y}^{CES}_{d,t} - \hat{Y}^{Respondents}_{d,t}}{\hat{Y}^{Respondents}_{d,t}},$$

relative overall error

$$RelOE_{d,t} = 100\% \frac{\hat{Y}^{CES}_{d,t} - Y^{ES-202}_{d,t}}{Y^{ES-202}_{d,t}}.$$

There is no uniform pattern found in the direction and magnitude of the error components. None of the error components appears to be a dominant factor of error (Fig.2). The plotting symbol in Figures 2 and 3 is the two-letter postal code for the state. December of 1999 (the 18th point after the benchmark month) is used for the display.

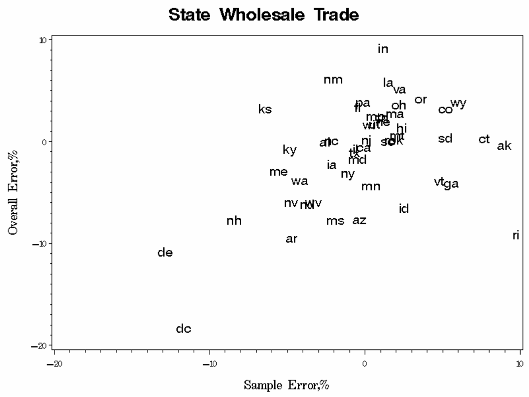

Figure 2. a) Relative Overall Error, $RelOE_{MID,t}$, vs. Relative Sample Error, $RelSE_{MID,t}$

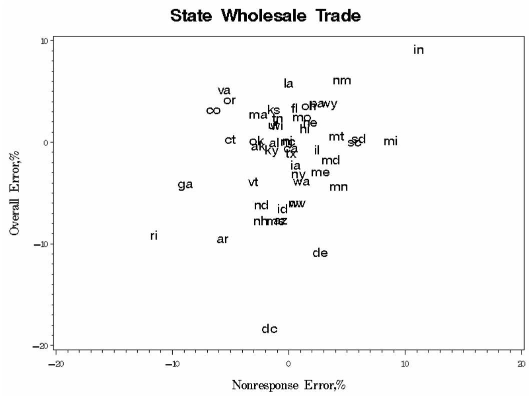

Figure 2. b) Relative Overall Error, $RelOE_{MID,t}$, vs. Relative Nonresponse Error, $RelNR_{MID,t}$

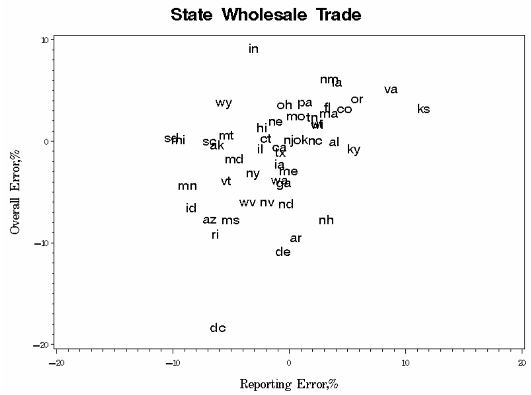

Figure 2. c) Relative Overall Error, $RelOE_{MID,t}$, vs. Relative Reporting Error, $RelRE_{MID,t}$

In general, low response rates would be of special concern if they were associated empirically with higher nonresponse bias. Figure 3 displays a plot of state-level relative absolute nonresponse errors against the corresponding weighted response rates for wholesale trade. Note that the plot does not display any pronounced pattern of association between nonresponse error and response rate.

To test the hypothesis about the presence of overall bias in the estimators $\hat{Y}^{CES}_{d,t}$, t values were computed as

$$t_{d,t} = \frac{\hat{Y}^{CES}_{d,t} - Y^{ES-202}_{d,t}}{\sqrt{V\hat{a}r[\hat{Y}^{CES}_{d,t}]}} .$$

The variance estimator $V\hat{a}r[\hat{Y}^{CES}_{d,t}]$ was computed by applying the BRR methodology to 6 strata in domain d.

In Figure 4, the t values are plotted against the quantiles of t distribution with 6 degrees of freedom. The serious negative bias may be attributed to various sources. For example, in Arizona ($t = -2.76$) and Mississippi ($t = -2.63$) it is dominated by reporting error. In Delaware ($t = -2.98$) and Washington State ($t = -2.75$) it is dominated by sampling error. In the District of Columbia ($t = -5.26$) it is dominated by combination of sampling error and reporting error.

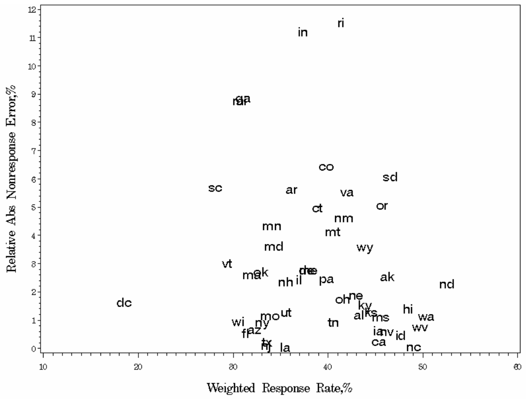

Figure 3. Relative Absolute Nonresponse Error, $\left|RelNR_{MID,t}\right|$, vs. Weighted Response Rate for State-Level estimates in Wholesale Trade

(It should be noted, however, that the estimates of variances used in computing of the t values may be unstable, especially in the domains with a low number of responding units. For example, the District of Columbia has only six responding units in wholesale trade in December 1999).

Overall, the deviations from the true values in the direct estimates $\hat{Y}^{CES}_{d,t}$ lay within respective confidence intervals.

3. Bias of the Synthetic Estimator

Let us analyze properties of the error components of the synthetic estimator $\hat{Y}^{CES,Synthetic}_{d=MID*MSA,t}$.

Note that algebraically,

$$RelNR^{Synthetic}_{MID*MSA,t} = RelNR^{Direct}_{MID,t}$$

and

$$RelRE^{Synthetic}_{MID*MSA,t} = RelRE^{Direct}_{MID,t}.$$

However,

$$RelSE^{Synthetic}_{MID*MSA,t} = \frac{R^{ES\,202}_{MID,t}}{R^{ES\,202}_{MID*MSA,t}} RelSE^{Direct}_{MID,t} + 100\%\left(\frac{R^{ES\,202}_{MID,t}}{R^{ES\,202}_{MID*MSA,t}} - 1\right)$$

where $R^{ES\,202}_{MID,t} = Y^{ES\,202}_{MID,t} \big/ Y^{ES\,202}_{MID,0}$

and

$$R^{ES\,202}_{MID*MSA,t} = Y^{ES202}_{MID*MSA,t} \big/ Y^{ES\,202}_{MID*MSA,0}$$

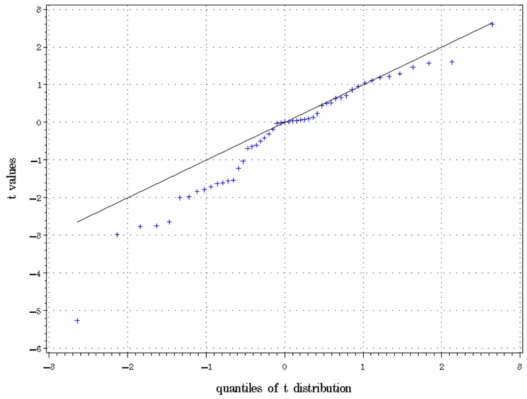

Figure 4. Ouantile-Quantile Plot of $t_{MID,t}$ Values Against the t_6 Distribution for State-Level Estimates in Wholesale Trade

are employment growth rates from benchmark month to month *t* at, respectively, state-level MID and MID*MSA level.

Therefore, the synthetic estimator $\hat{Y}^{CES,Synthetic}_{MID*MSA,t}$ is potentially subject to a bias due to the heterogeneity of employment growth rates across MSA within a particular industrial division. Consequently, it is important to evaluate the degree of this potential bias compared to other error components.

Let us define month t relative difference in the ES-202 growth rates of a state-level MID and MID*MSA intersection as:

$$RelR^{ES202}_{MID,MSA,t} = \frac{R^{ES202}_{MID,t}}{R^{ES\,202}_{MID*MSA,t}}$$

There is a strong positive correlation between the relative overall error of the synthetic estimator $\hat{Y}^{CES,Synthetic}_{MID*MSA,t}$ and relative difference in ES-202 employment growth rates $RelR^{ES\,202}_{MID,MSA,t}$ (Fig.5).

Finally, Figure 6 displays a quantile-quantile plot of the overall error in the synthetic estimator, $t^{Synthetic}_{MID*MSA,t} = \dfrac{\hat{Y}^{CES,Synthetic}_{MID*MSA,t} - Y^{ES-202}_{MID*MSA,t}}{\sqrt{V\hat{a}r[\hat{Y}^{CES,Synthetic}_{MID*MSA,t}]}}$ against a t distribution on six degrees of freedom. The pronounced deviations of the upper and lower tails of the $t^{Synthetic}_{MID*MSA,t}$ distribution from those of the t_6 distribution are consistent with nontrivial bias in the synthetic estimator.

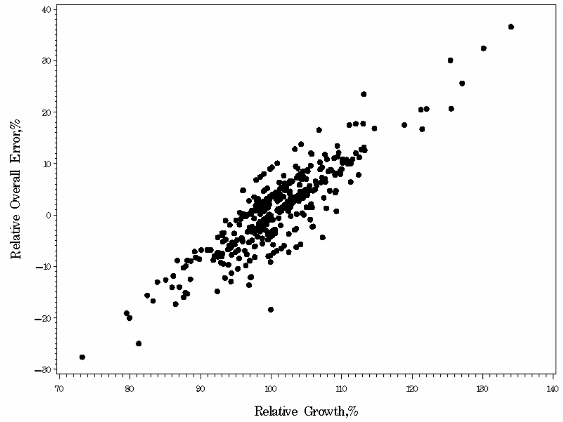

Figure 5. Relative Overall Error $RelOE^{Synthetic}_{MID*MSA,t}$ of the Synthetic Estimator vs. Relative Difference in Growth Rates $RelR^{ES202}_{MID,MSA,t}$

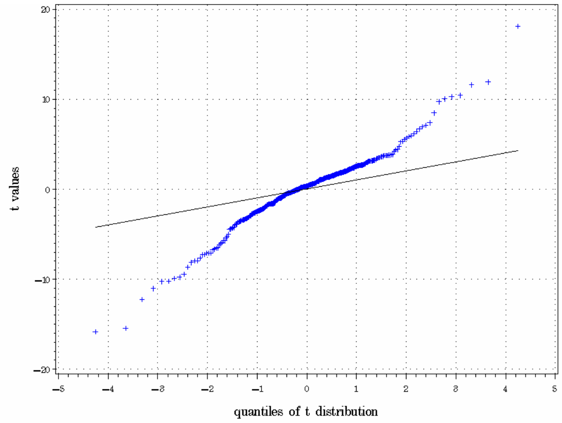

Figure 6. Quantile-Quantile Plot of $t^{Synthetic}_{MID*MSA,t}$ Values Against the t_6 Distribution

4. Conclusions and Further Research

There was no uniform pattern found in the sources of error of the direct sample estimates. The error may be attributed to sampling error, nonresponse effects, or differences in the data reported to the CES and to the ES-202.

The heterogeneity of MSA employment growth rates within industry leads to a bias of the synthetic estimator $\hat{Y}^{CES,Synthetic}_{MID*MSA,t}$. It is, therefore, important to make appropriate adjustments to it. This may be done by dividing a state into several homogeneous areas or by deriving a bias adjustment factor from available ES-202 information. For example, we can try to predict $RelR^{ES202}_{MID,MSA,t}$ from historical ES-202 data, and make an adjustment to the synthetic estimator:

$$\hat{Y}^{CES,Synthetic,Adjusted}_{MID*MSA,t} = \frac{\hat{Y}^{CES,Synthetic}_{MID*MSA,t}}{Rel\hat{R}^{ES202}_{MID,MSA,t}}$$

References

Butani, S., Harter, R. and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 523-528.

Butani, S., Stamas, G. and Brick, M. (1997). Sample Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Cochran, W.G. (1977). *Sampling Techniques,Third Edition.* New York: Wiley.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science* **9**, 55-93.

U.S. Bureau of Labor Statistics (1997). *BLS Handbook of Methods*. U.S. Department of Labor, Bureau of Labor Statistics Bulletin 2490, April, 1997. Washington, DC: U.S. Government Printing Office.

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

Wolter, K., Huff, L. and Shao, J. (1998) Variance estimation for the Current Employment Statistics survey. Presented at the Joint Statistical Meetings, Dallas, August 13, 1998.