SMALL DOMAIN ESTIMATION OF EMPLOYMENT USING CES AND ES202 DATA

Rachel Harter and Kirk Wolter
National Opinion Research Center
Michael Macaluso
Illinois Department of Employment Security

Abstract

The Current Employment Statistics program of the U.S. Bureau of Labor Statistics produces monthly estimates of employment, hours, and earnings at the national, state, and major MSA levels. Many states use the CES sample data to produce estimates of employment for local labor market areas, as well. The statistical redesign of the CES program introduced the idea of using ES202 employment as an auxiliary variable in estimation, not just in benchmarking and sample selection. Using ES202 data with simple model-based small area estimators can improve the labor market employment estimates, as demonstrated in simulations with CES and ES202 data for Illinois.

1. Introduction

The Current Employment Statistics (CES) program, a federal-state cooperative program of the Bureau of Labor Statistics (BLS), provides richly detailed and nearly real-time current data on nonagricultural employment, worker hours, and worker earnings by industry at the national, state, and metropolitan statistical area (MSA) level. Historically, the program began before probability sampling methods were well-developed. For the past few years, the BLS has been engaged in a research program with a committee of states and statistical consultants to revise and update the CES sample design, enrollment of sample in accord with the design, and estimation procedures. Much of the redesign research has been documented by Werking (1997), Butani, Stamas, and Brick (1997), Butani, Harter, and Wolter (1997), West et al. (1997), and Wolter et al. (1998). A key component of the research is the use of ES202 auxiliary data.

The ES202 program is another of BLS' federal-state cooperative programs. Each state collects employment data for virtually all employers in the state's unemployment insurance (UI) program. The data are not as current–lagging by six months–as the CES data, but they are available on a universe basis. Once per year, the CES is adjusted, or *benchmarked*, to conform to ES202 data, subject to differences in coverage and scope.

Illinois and other states produce estimates of employment at more detailed levels of geography than the CES sample was originally designed to support. In such cases statisticians often turn to *small domain estimation* techniques. The geographical dimension of Illinois' small domains consists of labor market areas (LMAs) outside of the MSAs already estimated by CES. Illinois' 56 LMAs are individual counties or small aggregations of adjacent counties with a common workforce and economy. The LMA estimates are receiving more attention since Congress passed the Workforce Investment Act of 1998 requiring the implementation of a system of national, state, and *local* employment statistics.

NORC and the Illinois Department of Employment Security (IDES) have conducted a joint research project to develop a method for producing employment estimates at the LMA level. The research has included a number of simulations using Illinois data to test several small domain estimators. Section 2 gives details on some of the simulations and estimators. Simulation results are presented in Section 3. Conclusions and recommendations for further research are presented in Section 4.

2. Small Domain Simulations Using CES and ES202 Data

2.1 Simulation Basics

The first step in the research project was to clearly define the scope of the problem. For this project, employment estimates were requested for 56 LMAs for 13 industry divisions, generally defined by 1-digit Standard Industrial Classification (SIC) codes.

To investigate the suitability of ES202 employment as an auxiliary variable, we reviewed the differences in coverage and scope between the CES and ES202 measures of employment. We found correlations between the two to be quite high, typically greater than .9. We reviewed scatter plots of CES employment with ES202 employment at the establishment level. We reviewed differences in editing procedures, processing systems, and schedules. In spite of all the differences, the ES202 employment, regardless of the time period, appears to be an excellent auxiliary variable for estimation with CES sample data, *provided CES and ES202 employment can be accurately linked at the establishment level*. For our small domain estimators, we chose to use March ES202 employment because it is readily available through the benchmarking process and is most likely to be used in CES estimation for larger domains.

Most of the well-known small domain methods are summarized by Purcell and Kish (1980), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994). Some of the estimators that seemed reasonable for our application were tested by a series of simulations of increasing complexity to estimate employment level and change. The simulations described here used actual CES sample reporters as the "universe" to be estimated. The term "reporters" is used to indicate that the sample did not consist only of individual establishments. Some reporters were entire UI accounts within a state, and some were county-level aggregations of establishments within a UI account. The reporters were matched to their counterparts in the previous March ES202 database. Nonmatches were excluded from the simulations. Similarly, reporters with incomplete data for the months of the simulations were also excluded. The remaining "universe" was a very stable subset of reporters from the CES sample. The number of reporters in the data file depended on the number of periods in the simulation.

We selected 100 samples from our "universe" to estimate employment. Because our universe in these simulations was itself a sample, subsamples were too small to realistically attempt estimation for the 56 LMAs in Illinois. Instead, we used the subsamples to estimate employment for the nine CES MSAs and three balance of state (BOS) regions, treating the MSAs and BOS regions as though they were LMAs. Subsample estimates using small domain estimators were compared with the full CES sample totals for the MSAs and BOS regions. We used all industries to estimate employment level for selected months, and three industries (manufacturing, finance/insurance/real estate, and services) to estimate five consecutive months (Sept. 1995 - January 1996) for employment change.

In selecting subsamples, we assigned all reporters to one of four size class strata based on March ES202 employment. We selected random samples of 10% from each of the strata except the stratum of largest firms, where we selected all firms with certainty. For every sample, MSA employment was estimated using several different small domain estimators. A brief description of the estimators is given below, preceded by some notation.

2.2 Notation

The notation below is written for estimation at the LMA level. For the simulations described here, MSAs and BOS regions were treated as LMAs.

i = 1, . . ., 13, the industry division

m = 1, . . . , the statewide industry model cell (see section 2.5)

ℓ = 1, . . ., the LMA (MSA in these simulations)

j = 1, . . . , the establishment

t = 0, 1, . . . , the month to be estimated (t = 0 for base period)

y_{jt} = CES employment for establishment j at time t

x_{j0} = ES202 employment for establishment j available at time t=0; base period ES202 employment

U_m = universe of establishments in model cell m (noncertainty universe for estimators 3-9)

s_m = sample of establishments in model cell m (noncertainty sample for estimators 3-9)

$D_{i\ell}$ = domain of industry i in LMA ℓ

$\delta_j (D_{i\ell})$ = 1 if establishment j is in $D_{i\ell}$
= 0 otherwise

$\hat{\beta}_m$ = estimated generalized regression parameter for model cell m

w_j = sampling weight for establishment j (1 for certainty establishments; 10 for noncertainty)

$Y_t (D_{i\ell})$ = total CES employment for domain $D_{i\ell}$ at time t = $\sum_{m \in i} \sum_{j \in U_m} y_{jt}\ \delta_j\ (D_{i\ell})$

$X_0 (D_{i\ell})$ = total ES202 employment for domain $D_{i\ell}$ for time t=0 = $\sum_{m \in i} \sum_{j \in U_m} x_{j0}\ \delta_j\ (D_{i\ell})$

$N (D_{i\ell})$ = the number of noncertainty establishments in $D_{i\ell}$ = $\sum_{m \in i} \sum_{j \in U_m} \delta_j\ (D_{i\ell})$

$\hat{N} (D_{i\ell})$ = an estimate of $N (D_{i\ell})$ = $\sum_{m \in i} \sum_{j \in s_m} w_j\ \delta_j\ (D_{i\ell})$

Note that $\sum_{m \in i}$ means summation over all model cells m that cross or intersect industry division i.

2.3 The Estimators

The most basic estimator in the simulations was the *simple unbiased estimator*, or the Horvitz-Thompson estimator, which uses only sample data within a small domain to estimate total employment for the domain. Any reasonable estimator we select ought to perform at least as well as the simple unbiased estimator (EST1), defined by

$$\hat{Y}_t\ (D_{i\ell}) = \sum_{m \in i} \sum_{j \in s_m} y_{jt}\, w_j\, \delta_j\ (D_{i\ell}), \quad \text{if } D_{i\ell} \text{ contains sample,}$$
$$= 0 \qquad\qquad\qquad\qquad\qquad\qquad , \text{ otherwise.}$$

The second estimator is a form of the *link relative estimator* [Madow and Madow (1978) and West (1983, 1984)], the historical mainstay of CES employment estimation. The link relative estimator (EST2) projects forward the prior month's estimate based on the trend among sample units that reported both months. In the simulations, the sample data in the trend component were weighted. We used the March ES202 employment for the small domain "universe" as the starting point for the link relative series. For a constant universe and sample, the link relative estimator simplifies to
We included a few *regression-based estimators* that incorporated the ES202 employment as auxiliary data. The class of regression-based estimators includes both the generalized regression and the "shrinkage" generalized regression estimators for small samples. Defined broadly, this class encompasses many of the estimators proposed in recent years, including synthetic estimators, sample regression estimators, Bayes estimators, empirical Bayes estimators, hierarchical Bayes estimators, and others. For more information on these estimators, see Datta and Ghosh (1991), Efron and Morris (1973), Fay and Herriot (1979), Gonzalez (1973), Hidiroglou et al. (1995), Holt et al. (1979), Hulting and Harville (1991), and Laake (1978). The regression-based estimators "borrow strength" from other domains by combining data from neighboring domains to reduce the variability in the estimates, assuming that the neighboring domains have similar data relationships.

In the simulations, the sample CES data were combined with predicted CES values for nonsample units, where the predictions were derived from linear models fit using the sample CES and ES202 data. Most predicted values were based on the same simple model. Ordinarily the model cells for estimating model parameters would be statewide 2-digit SIC groupings, but for these simulations on the reduced CES/ES202 matched file, the model cells were statewide industry divisions. Estimating the models statewide assumes that the model relationship is appropriate for all smaller geographies. The class of estimators is defined by

$$\hat{Y}_t\ (D_{i\ell}) = \sum_{certainty} y_{jt}\, \delta_j (D_{i\ell}) + \sum_{m \in i} \left[\sum_{j \in U_m} x_{j0}\, \hat{\beta}_m\, \delta_j\ (D_{i\ell}) + \lambda \sum_{j \in s_m} (y_{jt} - x_{j0}\, \hat{\beta}_m)\, w_j\, \delta_j\ (D_{i\ell}) \right],$$

where λ takes a value in the interval [0,1]. We tested the following values of λ.

EST3: $\lambda = 1$, the generalized regression estimator

EST4: $\lambda = 0$, the synthetic estimator

EST5: the Särndal-Hidiroglou (1989) shrinkage estimator

$$\lambda = 1 \qquad , \quad \text{if} \quad \sum_{m \in i} \sum_{j \in s_m} \delta_j \, (D_{i\ell}) \geq 20$$

$$= \; (\hat{N} \, (D_{i\ell}) / N \, (D_{i\ell}))^{H-1}, \quad \text{otherwise,}$$

where

$$H = 0, \text{ if } \hat{N} \, (D_{i\ell}) \geq N \, (D_{i\ell}) \, ,$$
$$= 2, \text{ if } \hat{N} \, (D_{i\ell}) < N \, (D_{i\ell}) \, .$$

For small domains with no sample data, the regression estimators reduce to

$$\hat{Y}_t \, (D_{i\ell}) = \sum_{certainty} y_j \, \delta_j \, (D_{i\ell}) \; + \sum_{m \in i} \sum_{j \in U_m} x_{j0} \, \hat{\beta}_m \, \delta_j \, (D_{i\ell}) \quad .$$

Another shrinkage estimator, a variation on Battese et al. (1988) was also tried. For this estimator, λ was a function of estimated variance components under a mixed linear model [Harville (1976), Henderson (1975), Robinson (1991)], resulting in an estimated best linear unbiased predictor under the model. As expected, we found very little difference among the regression-based estimators, so for most simulations we restricted our attention to simpler versions that did not require intermediate calculations of variance components or other quantities.

In a variation of the synthetic estimator (EST6), each sample unit represents itself, whether certainty or noncertainty. Each noncertainty unit is predicted using the model developed on the noncertainty sample units.

$$\hat{Y}_t \, (D_{i\ell}) = \sum_{certainty} y_{jt} \, \delta_j \, (D_{i\ell}) \; + \sum_{m \in i} \sum_{j \in s_m} y_{jt} \, \delta_j \, (D_{i\ell}) \; + \sum_{m \in i} \sum_{j \notin s_m} x_{j0} \, \hat{\beta}_m \, \delta_j \, (D_{i\ell})$$

A similar estimator (EST7) uses a different correction term for the model-based synthetic estimator. This estimator is unique because, in the absence of certainty units, it estimates the quantity $X_0 \, (D_{i\ell}) + Y_t \, (D_{i\ell}) - Y_0 \, (D_{i\ell})$. If ES202 and CES really are measuring the same definition of employment, then this estimator should be a suitable estimator of $Y_t \, (D_{i\ell})$. Otherwise, if they are merely highly correlated but different, then the estimator is a hybrid of an estimator of $Y_t \, (D_{i\ell})$ and an estimator of $X_t \, (D_{i\ell})$.

The simulations included two versions of *raking*, or iterative proportional fitting. Bousefield (1977) was among the first to describe the use of raking to force the marginal totals of a two-way sample table to match census totals. Raking is a special case of the "structure preserving" small domain estimators proposed by Purcell and Kish (1980). Raking for small domain estimation requires a starting value for each small domain. Then the small domain estimates are adjusted by ratios of known marginal totals to the sum of the estimates. A multi-dimensional table of small domain estimates can be iteratively raked to the corresponding sets of known marginals until the table converges. In this way the small domain estimates are forced to sum to all the marginal totals. The two raking variations tested in the simulations differed in the choice of starting values and marginal

totals. Only the noncertainty component was involved in the raking. Then the certainty units were added back at the end.

The raked estimator similar to that originally proposed by the BLS for MSA estimation (EST8) began with the simple unbiased estimator of ES202 employment in each cell for the base period. We iteratively adjusted the cell estimates to sum to the benchmark ES202 marginal totals for industry divisions and BOS geographies. The ratios between the converged cell figures and the initial cell values formed adjustment factors, which were then applied to the simple unbiased estimates of CES employment (EST1) for the estimation month. This estimator is zero when a small domain has no sample data.

An alternative raked estimator (EST9) started with a synthetic estimator (EST4) for the small domain at time *t*. The iterative raking took place at month *t* using generalized regression estimates of employment at higher levels of aggregation as marginal totals.

3. Simulation Results

To evaluate the estimators across the 100 samples selected, we computed biases and mean squared errors as mean differences or mean squared differences between the estimates and the "true" totals. Biases and root mean squared errors were also computed on a relative basis because the differences varied dramatically between Chicago and the remaining MSAs. Biases, relative biases, root mean squared errors, and relative root mean squared errors were averaged across industries and across MSAs for simple summary performance measures. See Figure 1 for average biases and root mean squared errors of the estimators of employment level. For simulations over time, we also averaged the summary statistics across months, as in Figure 2. Relative statistics were not computed for employment change because of problems with division by zero in periods of no change.

In general, the estimators that incorporated ES202 auxiliary data performed better than those that relied only on sample data within each small domain independently. The performance of the raked estimators varied greatly, depending on the initial estimates and the way in which estimates were produced for small domains with no sample data. The generalized regression and shrinkage estimators occasionally produced negative estimates due to the correction for lack of fit. Two versions of the synthetic estimator (EST 4 and EST 6 in the figures) performed best, on average, at estimating employment level as well as month-to-month change.

Figure 1 - Avg Bias and RMSE By Estimator of Level

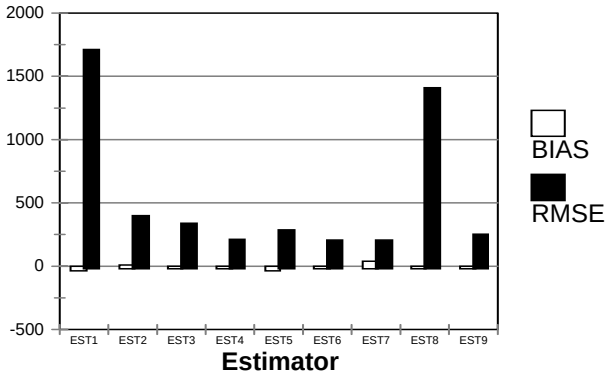


Figure 2 - Avg Bias and RMSE By Estimator of Change

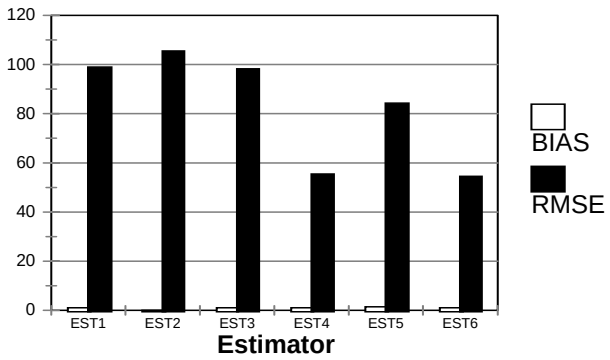


Table 1
Percentages of Change Estimates With Incorrect Sign
Manufacturing, FIRE, and Services Only

	EST1	EST2	EST3	EST4	EST5	EST6	EST7
Dec-Jan	15%	15%	19%	11%	18%	11%	11%
Total	13%	13%	18%	17%	17%	16%	17%

In estimating over-the-month change, it is important to get the direction of the change correct. Table 1 shows the percentages of December-January changes and total estimated changes (out of 30,000) with incorrect sign for the various estimators, excluding the raked estimators. While the percentages are at first a little disheartening, it should be noted that an estimator is more likely to have the incorrect sign when both the true change and the estimated change are not significantly different from zero. The link relative estimator (EST2) and the simple unbiased estimator (EST1) have the best overall record in Table 1, but the synthetic estimator EST6 is next best by this measure. Actually, EST1 and EST2 have more "no change" estimates because of the way estimates are handled in the absence of sample data. If any incorrect "no change" estimates were counted as incorrect signs, EST2 had 31% with incorrect sign, while EST6 had 17%. In Table 1 for the December-January period, which may be the toughest change to estimate, the synthetic estimators (EST4 and EST6) have fewer incorrect signs than the link relative estimator, even without counting "no change" errors.

In another comparison of the synthetic estimator with the link relative estimator over time, we computed empirical confidence intervals by plotting the 5^{th} and 95^{th} of the 100 sample estimates in order of magnitude for each time period. Within these bounds, we also plotted the median estimate and the true values. See Figures 3-4 for typical examples of plots of employment level and Figures 5-6 for examples of plots of employment change. The link relative (EST2) estimator was selected because of its familiarity to CES program staff. The synthetic estimator (EST4) and its variant (EST6), which were virtually indistinguishable in performance, were the best estimators overall. While results varied considerably from one small domain to another, the plots demonstrated rather

dramatically that the synthetic estimator typically outperformed the link relative estimator in estimating both level and change.

Figure 3 - EST2 Estimates of Level
FIRE - Bloomington/Normal

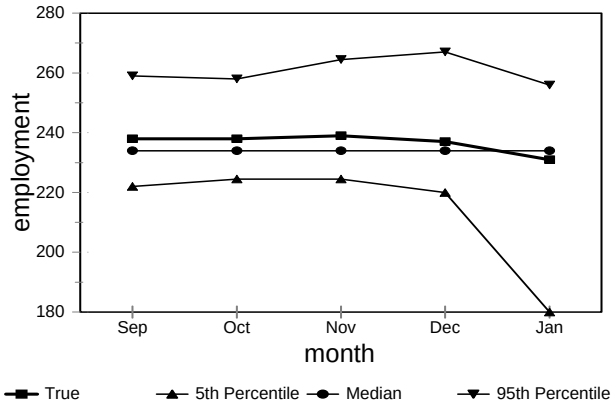


Figure 4 - EST4 Estimates of Level
FIRE - Bloomington/Normal

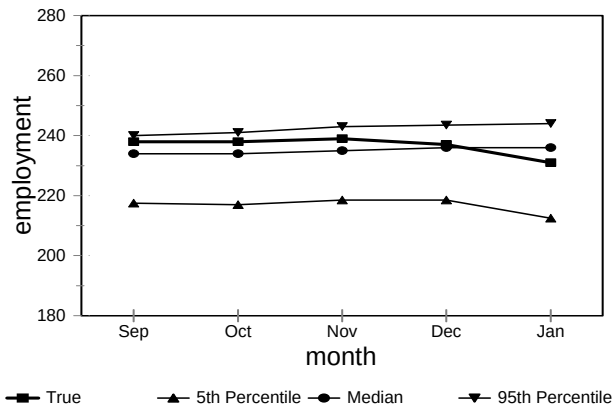


Figure 5 - EST2 Estimates of Change
FIRE- Bloomington/Normal

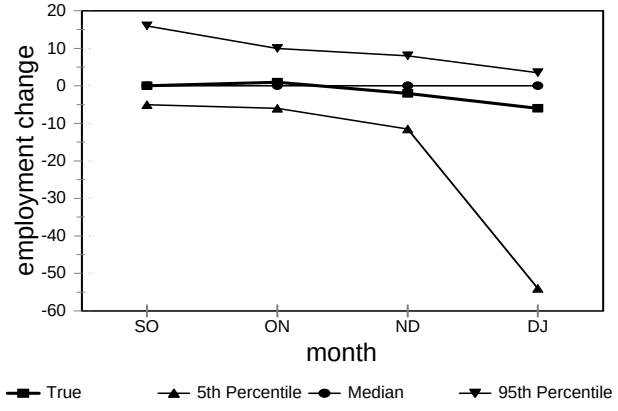
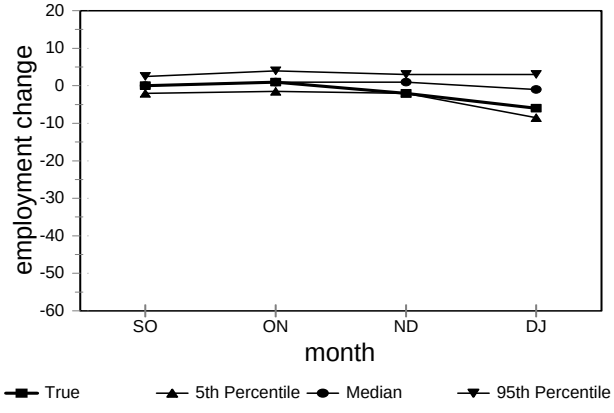


Figure 6 - EST4 Estimates of Change
FIRE- Bloomington/Normal



A well-known property of the synthetic estimator is a risk of bias when the model estimated at the state level is not appropriate for individual small domains. We are exploring ways in which we can tap into the knowledge of local labor market economists to keep the models from going too far wrong in any individual LMA.

Further simulations using the matched data file and the larger ES202 file tested variations of the synthetic estimator. For example, adding a simple one-dimensional rake to force small domain estimates to sum to estimates of higher level aggregations did not greatly affect the performance. We also tested the effectiveness of fitting the models at sub-state levels rather than at the state level. The sub-state model reduced the error in some difficult-to-estimate small domains, but not sufficiently to offset the increased variability in the models and the resulting errors overall.

4. Conclusions and Additional Research

Using ES202 employment as an auxiliary variable along with CES sample data improves the estimation of employment for small domains as applied to Illinois data. Variations of the synthetic estimator performed best, and EST6 was selected for further testing. Based on our tests so far, EST6 remains our best choice for estimating LMA employment in Illinois.

Illinois and NORC are building a small domain estimation "engine" and system for producing synthetic estimates for "building-block" domains. It is our intention that, by estimating employment for the smallest domains likely to be needed, the system will be sufficiently flexible to meet the requirements of Illinois' programs. If the system is sufficiently modular and portable, other states and other agencies may benefit from the development.

The timely incorporation of business births and deaths remains a challenge. Approaches that work well at aggregate levels may perform unsatisfactorily in smaller domains. We continue to investigate ways to improve small domain estimation in the context of births and deaths.

5. References

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Bousefield, M. (1977). Intercensal estimation using a current sample and census data. *Public Data Use*, **5**, 6-15.

Butani, S., Harter, R., and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics' Current Employment Statistics Program. *1997 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. *1997 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 517-522.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics*, **19**, 1748-1770.

Efron, B. And Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association*, **68**, 117-130.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Ghosh, M. And Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. *1973 Proceedings of the Social Statistics Section*, American Statistical Association, 33-36.

Harville, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, **4**, 384-395.

Henderson, C.R. (1975). Best linear unbiased estimation and predction under a selection model. *Biometrics*, **31**, 423-447.

Hidiroglou, M., Särndal, C.-E., and Binder, D.A. (1995). Weighting and Estimation in Business Surveys. In Business Survey Methods, eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A.Christianson, M.J. Colledge, and P.S. Kott. New York: John Wiley & Sons, 477-502.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, **74**, 405-410.

Hulting, F.L. and Harville, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association*, **86**, 557-568.

Laake, P. (1978). An evaluation of synthetic estimates of employment. *Scandinavian Journal of Statistics*, **5**, 57-60.

Madow, L. and Madow, W. (1978). On link relative estimators. *1978 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.

Purcell, N.J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, **48**, 3-18.

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, **6**, 15-51.

Särndal, C.-E. and Hidiroglou, M.A. (1989). Small domain estimation: A conditional analyses. *Journal of the American Statistical Association*, **84**, 266-275.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and strategies for small area data (with discussions). *Survey Methodology*, **20**, 3-22.

Werking, G.S. (1997). Overview of the CES (Current Employment Statistics) redesign. *1997 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 512-516.

West, S. (1983). A comparison of different ratio and regression type estimators for the total of a finite population. *1983 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 388-393.

West, S. (1984). A comparison of estimators for the variance of regression-type estimators in a finite population. *1984 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 170-175.

West, S., Kratzke, D., and Grden, P. (1997). Estimation for average hourly earnings and average weekly hours for the Current Employment Statistics Survey. *1997 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 529-534.

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.