DISCUSSION

Charles H. Alexander, Bureau of the Census
Demographic Statistical Methods Division, Washington, D. C. 20233

The Jackson and Frazier paper reminds us that even in the Age of Automation, expensive clerical work is still often required to merge multiple lists into a single frame. The need for clerical work is caused by lack of standardization of names and data fields for lists that were prepared with little or no coordination, often for disparate purposes. The desire to save money, and simplify processes so they can be more readily automated, creates a constant pressure to see whether multiple lists are really needed.

The main result of the paper is that multiple list frames are indeed needed to provide adequate coverage of the universe for the Private School Survey. Neither the State nor Association lists give adequate coverage alone. Even the lists for smaller Associations can have a noticeable effect on the coverage for affiliation categories. The Quality Education Data list does not seem to have added anything. However, this result needs to be double checked. The implied number of schools added to the QED between 1991 and 1993 is much smaller than seems reasonable; the procedure for identifying adds in this study should be reviewed carefully.

Can the results shed light on the completeness of coverage of the Private School Survey, using the existing frames? Perhaps, but more details would need to be recorded during the clerical operation, as described below.

Of course, when all the list frames included in the coverage study are actually used by the survey, it's impossible to prove any coverage deficiencies. For example, if a given State list did a very poor job covering the Association lists, this doesn't imply any undercoverage. After all, the Association-list Schools are covered by the Association lists and who's to say that the non-Association-list Schools aren't perfectly covered by the State list? Obviously this argument is dubious; such poor coverage of Association list would raise suspicions about the State list.

In this vein, we could seek indications of coverage problems by looking at the following.
For each Association and State:

i) what proportion of the Association-list Schools from that State are on the State list;

ii) what proportion of the State-list Schools with the relevant affiliation are on the Association list.

For States and Associations where these proportions are not high, questions should be asked about how the lists were put together to try to find out what's wrong. The second proportion is affected by inconsistencies in linking the affiliation information from the State list to the school's Association membership, as well as by the coverage errors we are looking for.

Although this information could be valuable, it would add steps to the clerical operation, so let me call this a suggestion rather than a recommendation, until the cost can be estimated.

The Kaufman, Li, and Scheuren paper gives a good illustration of the value of the Generalized Least Squares (GLS) methods of deriving survey weights, and the need for caution in using it. Their experience is similar to what was encountered in applying GLS to weighting for the Consumer Expenditure Surveys (Luery (1986), Zieschang (1986, 1990), Alexander (1987, 1990).)

Generalized least squares is a flexible, elegant method for making weighted estimates from surveys agree with as another or with controls derived from independent sources. But as the authors mention, it can have problems.

The most obvious problem, negative or very small weights, has several solutions. At a later session Jayasuriya and Valliant will present an appealing way of controlling the size of weights using the calibration estimation approach.

The more serious problem mentioned by the authors is the potential for harmful effects on estimates not directly controlled. We need a more complete theory of "harm" and "good" from the GLS method. The authors' "harm" measure is a step in the right direction.

At least part of the problem is that the "attractive asymptotic properties" of GLS do not apply when:

i) the survey has systematic undercoverage (Alexander, 1990); or

ii) the variables used to define the "control cells" have measurement error or are defined inconsistently between surveys; or

iii) as the authors note, when finite sampling properties apply, either because of a small total sample size or because of a few large sample units.

In these circumstances, the original weighted sample estimates may be very far from the controls, and the results can in fact be very sensitive to the "loss function" used. In the household weighting context, the loss function used by the authors responds to a large across-the-board undercoverage of households of all sizes by raising the weights of large households relative to small households. A different loss function increases all weights proportionally (Alexander, 1987, Table 1).

Kaufman, Li, and Scheuren propose a solution similar to what was ultimately used by the Bureau of Labor Statistics in applying GLS to the Consumer Expenditure Survey: adjust for "undercoverage " (or other systematic deviation from agreement with controls) before applying GLS to force agreement with controls. This is in effect what the Olkin method does. This makes sense on these assumption that the "attractive asymptotic properties" more nearly apply once this bias is reduced.

The authors are to be commended for looking hard at their data and not being awed by the elegance of GLS, nor frightened off by the need to use it carefully.

In his solo paper, Kaufman likewise looks closely at how new methods actually work for his data and his sample design. Kaufman proposes and implements a bootstrap variance method inspired by a discussion in Efron (1992). He has to extend Efron's treatment to handle the case of systematic sampling without replacement.

The paper describes an extensive evaluation via simulations based on real SASS and PSS data. As the author has explained, his method is to draw repeated samples and calculate confidence intervals from each sample, see what proportion of the intervals cover the simulated population parameter, and to compare these proportions to the nominal confidence level. The author's conclusion is that the bootstrap method does better than the balanced half sample method previously used for the PSS as well as the SASS, with a few exceptions.

There is an obvious concern about the evaluation method as described. The bootstrap variance depends very much on the sort order applied prior to selection of the bootstrap sample. The optimal sort order is chosen as the one that given the best results looking at data from the same simulation on which it is evaluated; this may not be a fair evaluation. However, I suspect that this problem does not affect the basic result, because the range of sort orders actually need in the simulation is fairly limited, and because of results in Kaufman (1993) that show the bootstrap's superiority does not seem to be much affected by the exact ordering.

This problem aside, there are still some unanswered questions:

i) why is Kaufman's method occasionally not better than the balanced half-sample replication method? When does this occur?

ii) how does the bootstrap method compare to other improvements to the basic BHR method, such as variations on the stratified jack-knife, or Bob Fay's idea of giving partial weight to the "excluded half-sample." Intuitively, Kaufman's method has some of the same beneficial effects as these methods. Could this be the reason it beats the relatively crude BHR method used for SASS and PSS?

We need a more comprehensive theory of when and why these methods work best, and why.

Smith, Ghosh, and Chang boldly sail into tempestuous waters. The choice of survey periodicity is usually made based either on explicit but overly simplistic models, or on ad hoc intuitive attempts to consider the full range of concerns. Their paper is a skillful attack on this hard, controversial problem, of systematically representing the complexity of the periodicity choice. They explore some innovative approaches, though they do not reach a final conclusion.

I'm particularly appreciative of the complexity of this problem because of my recent involvement in similar problems related to the Census Bureau's so-called "Continuous Measurement" survey. We decided on an every-year (indeed every month) periodicity based on a much less careful analysis than that of these authors, but now we find we do need to take their kind of care with respect to the choice of how many years' data to use in small area estimates.

Among the authors' alternative ideas, there are many I like a lot, and few I would question.

Things I liked a lot:

- using ARIMA models to describe possible "population" values;

- consideration of methods for "short time series";

- the analogy with the inventory scheduling problem;
- the authors' awareness of the ambiguity of the notion "total resources are fixed."

Things I'm less enthusiastic about:

- the assumption that if the survey designer assumes a particular ARIMA model to evaluate the best periodicity, then data users will use forecasts from that model to analyze the data;
- waiting for "reliable cost estimates of all relevant cost elements", even for periodicities never encountered in practice;
- focussing only on the unconditional properties of the estimates.

Users will do as they please regardless of the designer's assumptions. In some applications, such allocating funds, it may make sense to project ahead to the current year if a good model is available. For other applications, users will prefer the last direct cross-sectional estimate.

It is very hard to speculate how the operation would be organized for periodicities that have never been used in practice, and what it would cost. We'd be fortunate to get plausible ranges for the cost.

As do most statisticians, the authors focus on the unconditional properties of estimates. Some statisticians would disagree with this, as would many politicians. If the realized recent values of the time series for a State are such that the State estimates are adversely affected by a particular periodicity for the next few years, it is little consolation to explain that their recent values are the product of a process for which that periodicity works well on average.

My general suggestion about this problem is that the conclusion must consider various possible combinations of: i) ARIMA models; ii) independent variables; iii) analyses and data uses; iv) sets of cost parameters; v) loss functions; vi) approaches to the evaluation.

It is not reasonable to wait for a single final answer to the questions "what is the world like" and "what are the important uses of the data". Instead the best periodicity should be calculated for various combinations of the above considerations. Then for each periodicity, a statement could be made of the assumptions and uses which it best supports. This would help in focussing on exactly what time series measurement problems or rankings of priorities must be addressed to make the decision about periodicity.

References

Alexander, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Alexander, C. H. (1990). Incorporating person estimates into household weighting using various models for coverage. *Proceedings of the 1990 Census Bureau Annual Research Conference*, 445-462.

Luery, D. M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association*, 325-330.

Zieschang, K. D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association, 64-71.*