DISCUSSION

Wesley L. Schaible, U.S. Bureau of Labor Statistics

The past few years have seen a renewed interest in small area estimation problems. The papers presented in this session will all be valuable additions to the growing body of literature on small area estimation methods. I will comment briefly on each paper in the order of their presentation. A general comment will be included in the discussion of the Ghosh-Meeden paper and their estimator will be used to support the claim that the small area estimation problem and the two-stage cluster sampling problem are essentially identical when conditional arguments are employed.

The Fay paper is motivated by the practical problem faced by the Bureau of the Census of how to produce State estimates of median income for four-person families. The present Census method consists of two steps. First, the high variance CPS State estimates are fitted by a linear regression using low variance independent variables from the previous Census and other sources. Median income estimates are then made by calculating a weighted combination of the regression estimate and the original CPS estimate.

Fay and Herriot (1979) describe some modifications to the regression approach and propose an empirical Bayes estimator. The current paper presents a multivariate extension of the Fay-Herriot method. By simultaneously estimating the variable of interest and another, highly correlated variable the method proposed in this paper produces a small area estimate which "borrows strength" not only from other small areas, but also, from the correlated variable.

The paper compares two methods of estimating the variance-covariance matrix of the prior distribution, one from the CPS sample data and one by fitting the model to Census data from the previous time period. In an empirical study where the 1980 census medians were available as criteria to evaluate estimates, the paper compares errors of the earlier regression method and those of the two new methods. Both new methods performed better than the regression method. The method of estimating the variance-covariance matrix using Census data was chosen as the preferred method.

The empirical evidence given in the paper seemed mixed and more discussion of why the Census data method was chosen over the CPS data method would have been useful. Results from Table 1 seemed to support the CPS data method. The average absolute error over all States of the CPS method was less than that of the Census data method (1.99 to 2.12). The CPS method also had the smaller absolute error in 24 States and the larger in 21 States (6 States had errors of the same magnitude).

The Bayes theory behind this estimator assumes the medians are normally distributed. I would expect that median income estimates from small samples might not be normally distributed. Do we have any idea whether or not this is true and, if so, any idea of the impact on results?

A series of medians over time can behave erratically when they are calculated from multimodal distributions. A multimodal distribution can occur naturally as well as when respondents, rather than reporting to the requested degree of accuaracy, round to a nearby, "convenient" figure, e.g., a multiple of $500 or $1,000. Functions of medians, e.g., percent change between two medians, can also behave in peculiar ways. The method by which the orginal medians are calculated is not mentioned. The use of a weighted combination of the original small area sample median and the regression estimate can be viewed as a smoothing technique. It would be of interest to know if additional smoothing was introduced in the estimation of the original medians.

Ghosh and Meeden present a Bayes estimator of the finite population mean of a small area that is a weighted combination of the small area sample mean and the mean of the prior distribution. Estimators are suggested for the mean of the prior distribution and the ratio of the variance of the sampling distribution to that of the prior distribution in order to produce an empirical Bayes estimator. The authors show that their estimator for the inverse of the ratio of the variances is consistent and compare the Bayes risk performance of the empirical Bayes estimator with that of the "small area" sample mean and the "overall" sample mean. This is accomplished by employing the concept of relative savings loss. They also show that as the number of small areas approaches infinity the difference between the Bayes risk of the empirical Bayes estimator and that of the Bayes estimator approaches zero.

I'd like to describe some similarities between the small area estimation problem and the two-stage cluster sampling problem. In both problems our finite population is divided into clusters of units and we observe sample units within some but usually not all clusters. In both problems the clusters are almost always geographically defined although in neither problem is this necessary. On the surface, the objectives in the two problems seem different. The objective in the small area estimation problem is to estimate the population total (or mean) of each cluster of interest. The objective in the two-stage sampling problem is to estimate the population total (or mean) across all clusters, i.e., the sum of the cluster totals. The problem of estimating the

population total can be viewed as one of estimating each of the cluster totals and then summing them. If this view is taken then the small area problem and the two-stage cluster sampling problem are essentially the same. However, this line of thought necessitates the acceptance of theoretical arguments that are conditional on the selected sample. If we insist on a repeated sampling theory argument then an important distinguishing factor remains. That is, in the two-stage sampling problem we have control over the sample selection whereas in the small area estimation problem we must accept a sample that has been selected to meet other objectives. There seems to be almost unanimous agreement that conditional arguments are appropriate to address small area estimation problems. The appropriateness of conditioning for the two-stage cluster sampling problems is not so well accepted but the approach has a solid and growing constituency.

If we are willing to allow conditional arguments and accept the view that the two problems are essentially the same then we should expect results that apply to both problems. Under the model used by Ghosh and Meeden, but allowing the variance, τ^2, to vary between small areas, Scott and Smith (1969) consider the two-stage cluster sampling problem. They give an empirical Bayes estimator for the overall population mean within which is embedded an estimator for sampled cluster totals. This estimator can be written in the same form as the Ghosh-Meeden estimator. The suggested estimator for the mean of the prior distribution is the same. They suggest an estimator for the weight which is similar to the Ghosh-Meeden estimator in that it is also based on the F-ratio of between and within group mean sums of squares.

Royall (1976) approaches the two-stage cluster sampling problem from a non-Baysian prediction theory point of view. He gives the best linear unbiased estimator for a population total under a prediction theory model which specifies that the expected value of the variable of interest is a constant and that units within clusters can be correlated. This estimator contains an estimator for cluster totals which can be written in the same form as the Ghosh-Meeden and Scott-Smith estimators. Royall's estimator for the constant mean is similar to that in the other two papers, differing by the presence of covariance terms due to the addition of intracluster correlation to the model.

This model and others were used by Royall (1978) to investigate small area estimators in a paper presented at the Conference on Synthetic Estimates for Small Areas at Princeton. The similarities in results of the small area estimation and two-stage cluster sampling problems are apparent and I think work in each area might contribute to that in the other.

I'd like to make two observations on the Ghosh-Meeden paper from a practitioner's viewpoint. The practitioner is in the position of having to choose one method to use and being able to justify that choice. We have a number of methods from which to choose; many are generated by entirely different theoretical approaches. Practitioners need additional comparisons within the set of empirical Bayes methods and also with methods generated by other theoretical approaches. Comparisons of a given empirical Bayes estimator with the small area sample mean and the overall mean such as the ones in this paper are useful but are not adequate for the practitioner's needs. This comment can also be made for the other papers in this session and most, if not all, of the small area estimation literature. This is not an easy problem to address and it remains a major practical concern.

A second observation is that many of these models seem restrictive in that they assume the variable of interest is not expected to vary across small areas. It is difficult for a practitioner to justify selecting an estimator derived under the assumption that all small areas are expected to have the same mean value. This issue has been addressed to some extent in a paper by Ghosh and Lahiri (1986) in which stratification is introduced into an empirical Bayes argument.

The problem of suggesting an estimator for a small area estimation situation is usually involved and researchers often limit papers to the development of the estimator itself. Some papers take an additional step and present expressions for the variance or mean squared error. The Prasad-Rao paper is of particular interest in that it goes further and addresses the problem of estimating the mean squared error of an estimator for a given area. This is a difficult problem and one that deserves attention.

Prasad and Rao note that small area estimation models proposed by Battese-Fuller, Dempster et al., and Fay-Herriot are special cases of a general mixed linear model studied by Henderson in 1975. They also note that the three estimators of a small area mean associated with these models are best linear unbiased predictors (BLUPs). The BLUPs are functions of unknown variances. This paper uses the method of fitting constants and a jackknife procedure to estimate the unknown variances and produce estimated BLUPs. In addition, second order approximations to the mean squared errors of the estimated BLUPs are shown and both normality-based and jackknife estimators of the MSE's for the three models are derived. In an empirical study, relative efficiency comparisons are made between the Battese-Fuller BLUP estimated with fitted constants and two alternative estimators, the regression synthetic

estimator and the approximately design-unbiased regression estimator. The gains in efficiency obtained by using the estimated BLUP rather than the alternative estimators are impressive when the data are generated with normal errors. Estimating the BLUP with the jackknife procedure is slightly more efficient than with the method of fitted constants. The second order approximation to the MSE appears to work well especially when one or both errors in the model are normal or uniform. The relative biases of the two MSE estimators are small for normal and uniform errors and slightly larger for exponential and double exponential errors.

I have two brief comments on this paper. Empirical comparisons were made for the Battese-Fuller model; additional comparisons for the random regression coefficients model and the Fay-Herriot model would also be of interest.

In the empirical study the relative biases of normality-based and weighted jackknife estimators of the MSE both increased as the sample size in a small area increased. I found this puzzling and was curious as to whether or not we might expect this to continue as the sample size increases beyond that considered in this paper.

REFERENCES

Fay, R.E. and Herriot, R. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association, 74, 269-77.

Ghosh, M. and Lahiri, P. (1986), "Robust Empricial Bayes Estimation of Means from Stratified Samples," Technical Report No. 254, Department of Statistics, University of Florida.

Royall, R.M. (1976), "The Linear Least Squares Prediction Approach to Two-Stage Sampling," Journal of the American Statistical Association, 71, 657-64.

Royall, R.M. (1978), "Prediction Models in Small Area Estimation," Synthetic Estimates for Small Areas, NIDA Monograph 24, U.S. Government Printing Office, 63-87.

Scott, A. and Smith, T.M.F. (1969), "Estimation in Multi-Stage Surveys," Journal of the American Statistical Association, 64, 830-40.