Hyman B. Kaitz, C S R Associates

Both of these papers represent progress in the development of procedures for estimating local area data. Nevertheless, while they are work in progress, there is still some distance to go in attaining fully acceptable methodologies.

Both papers examine the accuracy of their techniques in terms of mean square errors of all the local area estimates, and in terms of the percentage of areas whose estimates deviate by more than a certain percentage, say 10%, from the known criterion values. It is desirable and relevant, of course, that such measures of dispersion or accuracy be used in judging these various techniques.

There is a particular problem, however, which arises in this connection. Estimates are produced for areas which are individually identified. For example, the method may yield a specific estimate for Altoona. Is Altoona interested in knowing whether its estimate comes from a body of area estimates with a satisfactory MSE, or is it interested in knowing how good the Altoona estimate is? I think the latter is more likely to be the case. In the absence of any alternative estimates. Altoona can perhaps become reconciled to its estimate, but it may not. It is interesting in this connection to see what has happened to local area population counts from the 1970 Census. Even though these are official government counts, they have been subjected to strong criticism by various local interest groups who have said that the presumed population undercount should be officially allocated to local areas differentially (blackinterest groups), or that the Census Bureau should use more accurate enumeration techniques (Spanish-American interest groups).

When the data at issue are statistical estimates rather than official counts there may even be more opportunity for criticism. For example, local area unemployment rate estimates have been subjected to criticism for a number of years in states like California, Ohio, New Jersey and Massachusetts, principally because these estimates, based on administrative records from the unemployment insurance system, could be compared with the rates for the same areas based on the Current Population Survey. The prospect for critical examination of local area estimates is particularly strong when concrete incentives are present to seek estimates which increase local allocation of government funds under programs such as revenue sharing.

This suggests an approach to local area estimation which is not based on the use of a single technique, but on all the information available for a given locality which may help to improve its estimate. The amount of such information may differ from one locality to another. This would be a highly professional-laborintensive activity and would probably be out of reach of most organizations with limited budgets.

With respect to the Erickson paper a question may be raised about the possible use of additional local area information available generally, which may be used as symptomatic variables, such as:

- 1) the racial mix in the base year
- 2) the urban-rural mix in the base year
- 3) the population density in the base year
- 4) The age-sex mix in the base year,
 - and so on.

Erickson's work uses ratios as the symptomatic variables as well as the criterion variable. I would like to see some evidence on the amount of collinearity present among the symptomatic variables. In general, on statistical grounds, one seeks for symptomatic variables which are relatively uncorrelated with each other but are correlated with the criterion variable. These considerations should not, however, ignore those based on the subject matter under study.

I like Erickson's formulation of the regression equation to take account of the sampling error in the criterion variable. This assumes independence of the u and v error terms which does not appear to be quite true for his data set. This suggests extension of his model to include some nonzero but unspecified covariance between u and v. In addition, there may be reason to assume a v term for specific symptomatic variables, which would alter the model somewhat and would appear to produce biased estimates of the parameters of the regression equation in the ordinary least squares approach. There is also some heterogeneity in u, khich suggests seeking some transformation of the variables to correct this, or the use of generalized least squares in the estimation procedure.

Between the extreme of seeking a single estimation equation for all areas and the extreme of seeking to maximize the use of ad hoc local area information through a variety of techniques, there may be an optimum point at which clusters of areas may be studied, each cluster with its own estimation technique. Erickson's formulation suggests little payoff here if the analysis uses sample-based criterion variables.

Erickson's focus is on change over a long period of time. He suggests that for shorter periods, but still multiples of years, estimation may become somewhat more uncertain, and may call for the use of fewer symptomatic variables. Gonzales is concerned with the use of synthetic estimates for shorter periods of time, such as months or quarters. Here additional questions may arise with respect to volatile variables such as unemployment. For example, a local area may experience a sizeable layoff at a large plant, which would introduce an evident discontinuity in its time series, a discontinuity which would not be reflected in the synthetic estimates tied to age-sex-color cells. I believe that the proposal to use occupation or industry cells which reflect changes in economic conditions more sensitively will do a better job, but that they will still reflect economic conditions in particular localities rather imperfectly.

When benchmark values are available, e.g. 1970 Census area estimates, it may be possible to apply a correction to the synthetic estimates. Let c_{i0} be the Census estimate for the i-th area, and x_{i0} be the synthetic estimate for the same area at that time. The difference, $d_i = c_{i0} - x_{i0}$, may then be taken as a first approximation adjustment for other time periods:

 $c_{it}^* = x_{it}^* + d_i$ where c_{it}^* is the adjusted synthetic estimate. I expect that such an adjustment would help considerably in improving the synthetic unemployment rate for Honolulu, for which the latter is quite deficient. The use of other corrections should be explored as well, perhaps some which are functions of time or specific variables.

Finally, it is possible that synthetic and regression techniques may be combined in some way as to yield a combination superior to either alone.