ORGANIZATION OF SMALL AREA ESTIMATORS

David A. Marker, Westat, Inc.

Introduction

Beginning in the late 1960's, the special problems of deriving estimates for small areas or domains from sample surveys have received increasing attention. Practically all attempts to derive such estimators have been either ad hoc approaches for specific problems or attempts to apply large-sample sampling theory to problems of small samples. Little emphasis has been placed on the assumptions underlying these methods or on the interrelationships among them.

Previous examinations of the various small area estimation techniques have merely listed the variety of techniques that are in use. In this paper I have coherently organized the different estimators, showing where certain methods can be viewed as trivializations or generalizations of others: some fall simply into a regression framework, others do not. From this organization a clearer understanding of the present techniques and their interrelationships will hopefully develop, as will an optimal path toward the further generalization of existing estimators.

Symptomatic Accounting Technique

The Symptomatic Accounting Technique (SAT) is a simple additive linear model for the change in the variable of interest. A typical example of the SAT (excluding migration for simplicity) is to assume that :

Population at time t+i = Population at time t + Births - Deaths.

This can be rewritten as a simple case of multivariate regression

$$\underline{Y} = X\underline{B} + \underline{E} \qquad \underline{E}\ N(\underline{0},\ {}^2I) \quad ,$$

Y_j = Population growth in area j from time t to time t+i,
$X = (\underline{X}_1,\underline{X}_2)$ = n x 2 design matrix,
$\underline{X}_{1j}$= Births in area j from time t to t+i,
$\underline{X}_{2j}$= Deaths in area j from time t to t+i,
$\underline{B} = (\underline{B}_1,\underline{B}_2)'$,
$j = 1,2,\ldots,n$,
$\underline{Y},\underline{X}_1,\underline{X}_2$, and $\underline{E}$ are n x 1 column vectors.

It is trivial to show that under an assumption such as that made by SAT, i.e.

$$Y_j = X_{1j} - X_{2j}$$

the least squares regression estimate, $\underline{B} = (X'X)^{-1}X'\underline{Y}$, based on these two symptomatic variables is:

$$\underline{B} = (B_1,B_2)' = (1,-1)' \ .$$

While this result is not particularly informative on its own, when viewed from this perspective the SAT is seen to be nothing more than a restricted case of the multivariate symptomatic regression method.

Vital Rates

Bogue (1950) suggested a generalization of the SAT known as the vital rates technique which uses the changes in birth and death rates rather than the raw values of the changes. This method assumes that the ratio of the birth (or death) rate for a given small area to the birth (or death) rate for the larger region remained constant since the last census. If such rates have been stable or falling since the last census then this assumption of a constant ratio may be a close approximation to reality. For example, if birth rates throughout the region had fallen by an average of ten percent, then a constant ratio would be preserved if the rate had fallen the most in those small areas with the highest birth rates and least in those areas that already had small rates. If however, the rates are rising, a constant ratio would require areas with high rates to grow even faster than before while areas with smaller than average rates would have to fall even farther behind the norm. To quote Bogue, "The small amount of specific evidence available indicates that the reverse of this assumption is true."

The vital rates estimate of the population total for local area 1 at time t is:

(1) $\quad Pop_{1t} = B_{1t}/BR_{1t}$

where B_{1t} = number of births in local area 1 from time t-1 to t,
BR_{1t} = birth rate in local area 1 from time t-1 to t.

The number of births, B_{1t}, is available from hospital or local records, the birth rate, BR_{1t}, however, must be estimated. The vital rates technique assumption of a constant ratio of birth rates over time

(2) $\quad BR_{1t}/BR_{st} = BR_{1t-1}/BR_{st-1}$

where BR_{st} = birth rate in state s (which contains locality 1) from time t-1 to t,

can be rewritten as

(3) $\quad BR_{1t} = (BR_{st}/BR_{st-1})BR_{1t-1}$.

This is how the vital rates technique estimates BR_{1t} and in turn Pop_{1t}.

The vital rates technique assumes that no knowledge is gained about state s or its localities by examining other states; therefore, it is appropriate to develop a separate univariate regression model for each state. One possible univariate regression model for each state would be

$$(4) \qquad Y_l = B\ X_l + e_l \quad E(e_l) = 0\ ,$$
$$V(e_l) = {}^2X_l/W_l\ ,$$

where $Y_l = BR_{lt}$ $\quad X_l = BR_{lt-1}$
and $W_l = N_l/N_s$.

This variance assumption has a logical basis in that it implies localities with large birth rates will be most variable as will those localities with smaller populations, N_l. If we make the additional assumption that the relative sizes of the different localities, N_l/N_s, have not changed from time t-1 to t, then the weighted least squares estimate of B is:

$$(5) \quad B = \frac{{}_{l\ s}\ Y_lX_l/(X_l/W_l)}{{}_{l\ s}\ X_l^2/(X_l/W_l)} = \frac{{}_{l\ s}\ W_lY_l}{{}_{l\ s}\ W_lX_l}$$

$$= BR_{st}/BR_{st-1}\ .$$

This leads to a regression estimate of

$$(6) \quad BR_{lt} = B\ X_l = (BR_{st}/BR_{st-1})BR_{lt-1}\ .$$

This is the same estimate found in equation (3). Therefore, under the model described by equation (4) the vital rates technique is simply the weighted least squares solution to univariate symptomatic regression.

Bogue suggests deriving two estimates of the population, one using birth rates and the other using death rates, and then averaging the estimates. As with the SAT this is just a special case of symptomatic regression, here taking the simple average of univariate regressions.

The United States Bureau of the Census (1974,1980) currently averages three different methods to derive its population and per capita income estimates for states, counties, and sub-county areas. These methods are multivariate symptomatic regression, administrative records, and component method II. The four symptomatic variables used in the regression equation for population are school enrollments, number of Federal Income Tax returns, car registration, and size of work force. Both administrative records and component method II are composite methods using symptomatic accounting techniques to track births, deaths, and the elderly and vital rates methodology to track migrations at the sub-county level. These are modern day applications of the composite method first suggested by Bogue and Duncan (1959). Just as each of these methods has been shown to be a form of regression, so to are the Bureau of the Census' combination methods.

Schmitt and Crosetti (1954) generalized Bogue's vital rates method into a multivariate framework. The univariate regression interpretation of vital rates shown in equations (3)-(6) can be similarly generalized to the multivariate setting.

Symptomatic Regression

The multivariate symptomatic regression method uses the model

$$\underline{Y} = X\ \underline{B} + \underline{E} \qquad \underline{E}\ \ N(\underline{0},\ )$$

where Y_i, i=1,2,...,N are the ratios for the dependent variable of the most recent census to the preceeding census, and X_{ij}, j=1,2,...,p are the ratios for the p symptomatic variables of the most recent census to the preceeding census, to derive least squares estimates, $\underline{B} = (X'X)^{-1}X'\underline{Y}$, for $\underline{B}$.

Then the estimates $\underline{Y}^* = X^*\ \underline{B}$ are found where X* are the ratios for the p symptomatic variables of their present values to their values at the time of the most recent census. This estimate of change Y_i^* is then multiplied by the most recent census value for the small area to give the symptomatic regression estimate of the present total for small area i. Multivariate regression on a set of symptomatic variables can therefore be seen as a general framework from which the symptomatic accounting technique, vital rates, and symptomatic regression technique are all special cases.

Sample Regression

One of the major drawbacks of the symptomatic regression technique is that it assumes a constant relationship between the independent and dependent variables reaching back to the census preceeding the most recent one. This is a sufficient but not necessary condition for the sample regression method introduced by Ericksen (1973). Ericksen's method also makes use of the most recently available data in the form of a sample from selected small areas. The sample regression method can be written as the following multiple regression equation:

$$\underline{Y} = X\ \underline{B} + \underline{E} \qquad \underline{E}\ \ N(\underline{0},\ )$$

where Y_i, i=1,2,...,n are the ratios of the sample estimate in area i to the previous census value for area i. X_{ij}, j=1,2,...,p are ratios for the p symptomatic variables of their present values to their previous census values in area i. The least squares estimates of $\underline{B}$ are then used along with the X_{ij}, to compute the estimates for all desired small areas. Instead of relying on the consistency of the relationship between the X and $\underline{Y}$ since the census before last, this only assumes that the relationship holds across all areas i since the last census. The sample regression method also depends upon the representativeness of the two-stage sampling process, choosing a sample of n areas and then sampling within those areas. This less restrictive set of assumptions clearly shows the sample regression method to be one more generalization in the regression approach to small area estimation.

Synthetic Estimation

Holt, Smith, and Tomberlin (1979) tried to show that using a prediction approach synthetic estimation can be placed in the general linear models framework with

(7) $$Y_{ijk} = B_j + e_{ijk}$$

where Y_{ijk} = kth element of small area i and subgroup j,
B_j = mean value for all of subgroup j,
e_{ijk} = error term distributed N(0, ^{2}),
i = 1,2,...,I,
j = 1,2,...,J, and
k = 1,2,...,N_{ij} .

This is the simple one-way analysis of variance model which makes the assumption, just like synthetic estimation, that the expected value of all elements belonging to a subgroup j will be equal, regardless of the small area it comes from. Under model (7) the lest linear unbiased estimate (BLUE) of B_j is

$$B_j = \bar{y}_{.j.} = {}_{i}\ {}_{k\ s}\ y_{ijk} / {}_{i}\ n_{ij}$$

where k s denotes those elements in the sample and n_{ij} denotes the number of sampled elements in small area i and subgroup j.

The BLUE of the average in small area i is given by

(8) $$\bar{Y}_{i..} = ({}_{j}\ n_{ij} (\bar{y}_{ij.} - \bar{y}_{.j.}) + {}_{j}\ N_{ij}\bar{y}_{.j.})/N_{i.} \quad .$$

This is not the synthetic estimate. The synthetic estimate is the last term in equation (8). Unless there are no sampled data from small area i (in which case the first term in (8) is equal to zero) the two estimators will disagree. Using the prediction approach, therefore, does not lead to the synthetic estimator except in an extreme case. Rather, the best linear unbiased estimate is to use the observed data directly and the synthetic estimate only to predict the nonsampled values.

Levy (1978) attempted to demonstrate how synthetic estimation can be put in a multivariate regression framework with the subgroup means being the unknowns estimated from the sample. The independent variables in this regression are then the P_{ij}, the population proportions within the small area. If each subgroup cross-classification (e.g., 30 to 40 year old black males) is viewed as a separate symptomatic variable then the synthetic estimator is seen as a linear combination of symptomatic variables:

$$\bar{Y}_{i..} = P_{i1}B_1 + P_{i2}B_2 + P_{i3}B_3 + \dots$$

In order to derive the standard synthetic estimate from such an equation requires each regression coefficient, B_j, to be equal to the sample average for subgroup cross-classification j. This is the least squares estimate only if instead of the population proportions P_{ij}, the independent variables were treated as indicator variables d_j whose value depended upon whether or not a given element was from subgroup j.

$$d_{j'} = \begin{matrix} 1 \text{ if } X_{ijk} \text{ is such that } j' = j \\ 0 \text{ otherwise.} \end{matrix}$$

An ordinary least squares regression equation of this form

$$Y_{ijk} = d_1B_1 + d_2B_2 + d_3B_3 + \dots$$

would result in the least squares estimate

$$y_{ijk} = {}^{J}_{j'=1}\ d_{j'}\ \bar{y}_{.j'.} = \bar{y}_{.j.} \quad .$$

The regression estimate of the average in small area i, $\bar{y}_{i..}$, would then be equal to the synthetic estimate:

$$\bar{y}_{i..} = \frac{1}{N_i}\ {}_{j,k}\ y_{ijk} = \frac{1}{N_i}\ {}^{J}_{j=1}\ N_{ij}\ \bar{y}_{.j.}$$

(9) $$= {}^{J}_{j=1}\ P_{ij}\ \bar{y}_{.j.} \quad .$$

Therefore we can only derive the synthetic estimate from ordinary least squares when we regress the y_{ijk} on a set of dummy variables, not directly on the population proportions (as Levy suggested). This shows the synthetic estimator to be equivalent to a form of the post-stratified estimator $y_{ijk} = \bar{y}_{.j.}$ often used in applied statistics (e.g., in imputation for nonresponse in a state with few respondents, imputing the regional average.) The derivation of the synthetic estimator from this regression equation depended upon using ordinary least squares. This homogeneity of the variances across subgroups is essential yet is unmentioned in any of the literature on synthetic estimation.

Synthetic estimation uses past census data only to determine the proportional decomposition of each small area into the different subgroups. It doesn't use any auxiliary variables and therfore doesn't make any assumptions about historical correlations among variables. The synthetic estimate's restrictive assumptions are that the average response for a given subgroup is the same in every small area and that the weights P_{ij} have remained constant since they were last measured. As a result of these assumptions, the synthetic estimator tends to underestimate large deviations from the overall average that occur in some small areas. Two adjustments to the standard synthetic estimator have been proposed to make it more sensitive to these deviations.

Composite Estimators

Schaible, Brock, and Schnack (1977) and Schaible (1978a,1978b) suggest a composite estimator using a weighted average of both the synthetic estimate and the simple inflation estimate

$$\bar{Y}_{i..} = \overset{J}{\underset{j=1}{}} N_{ij} \bar{y}_{ij.}/N_{i.}$$

from the sample where the weights are inversely proportional to the mean square errors of the two estimates. This method is now being used at the National Center for Health Statistics to calculate state level disability estimates. A problem with this method is that if no members of the sample come from a particular small area then its composite estimate is simply equal to the synthetic estimate.

Fay and Herriot (1979) show an excellent example of applying an empirical Bayes technique to the problem of small area estimation. Previously the United States Bureau of the Census had substituted county estimates for unknown values within a county. They proposed and have implemented a James-Stein estimation procedure to improve these estimates. They derive two separate estimates for the small area, a sample regression estimate based on past census symptomatic data, and a direct sample estimate based on the census' twenty percent sample in the small area. An average lack of fit for the regression model was calculated as was the sample variance for the twenty percent sample.Using the inverses of these variance estimates as weights, a combined James-Stein style estimator was used. In order not to deviate too far from the sample estimate the final estimate was constrained to be within one standard error of it. This estimator was empirically demonstrated to be superior to the previously used county estimates. This approach was modelled with the sample Y_i ind N($_i$,D) with $_i$ ind. N(A,V); D, A, and V known constants. The strength of a procedure such as this is in its flexibility, if some small areas appear to deviate further from the regression estimate than do others, you simply increase the weight given that sample estimate. This composite estimator is similar to that suggested by Schaible but here they have replaced the synthetic estimate with a regression estimate.

Synthetic Regression

The other alternative is to combine the synthetic estimate with the regression methods. Levy (1971) first suggested that the percent deviation of the synthetic estimator from its true value could be modelled as the dependent variable in a symptomatic regression equation. This can not be solved directly since the true values, and therfore the percent deviations, are unknown. He suggests estimating the regression coefficients based on collapsed strata of small areas and then using those values to revise the synthetic estimates for each small area. Recently Nicholls (1977) and Gonzalez and Hoza (1978) have taken the more direct approach of simply incorporating the synthetic estimator, along with the symptomatic variables, as independent variables in the sample regression method. These methods are known as synthetic regression.

Each of the estimators discussed before synthetic estimation involved regression on a set of continuous variables while synthetic estimation was shown to be regression on a set of categorical variables. Synthetic regression can therefore be viewed as a mixed continuous/categorical regression model.

Conclusion

This paper has described the interrelationships among the different small area estimation techniques. They can be viewed in an hierarchical structure where each method builds upon the others either by loosening the assumptions or including new sources of information.Table 1 summarizes and Figure 1 pictorially demonstrates this structure with each arrow pointing towards the more general type of estimator. Bogue (1950) generalized the SAT into the vital rates technique by using the rates of change since the previous census. We have demonstrated here how both the SAT and vital rates are special cases of multivariate regression on symptomatic variables, with vital rates using a less restrictive set of assumptions. Clearly, symptomatic regression is also a form of multiple regression.

We have shown how synthetic estimation can be seen as a form of symptomatic regression on categorical variables and its relationship to the commonly used post--stratified estimator. Ericksen (1973,1974) suggested generalizing the regression techniques to include sample data. This enabled him to loosen the assumptions necessary for symptomatic regression. This sample regression technique can be combined with synthetic estimation to develop the synthetic regression techniques of Levy (1971) or Nicholls (1977) or Gonzalez and Hoza (1978). The sample regression technique may also be combined with the direct sample estimate to derive the composite estimator suggested by Fay and Herriot (1979).

Combining the sample estimate with the synthetic estimate results in the composite estimator introduced by Schaible, Brock, and Schnack (1977) and Schaible (1978a, 1978b). Eventually we are left with just two building blocks: synthetic regression and composite estimators.

For many small areas there is often no sample available to use in a composite estimator. The synthetic regression estimator is always usable, but is only appropriate if a linear relationship exists among the available variables. Neither of these methods will be preferable 100 percent of the time. More work is necessary on balancing the relative advantages and disadvantages of the different methods in order to develop a consistent methodology for choosing the best small area estimator for any particular situation.

TABLE 1: Summary of Small Area Estimation Techniques

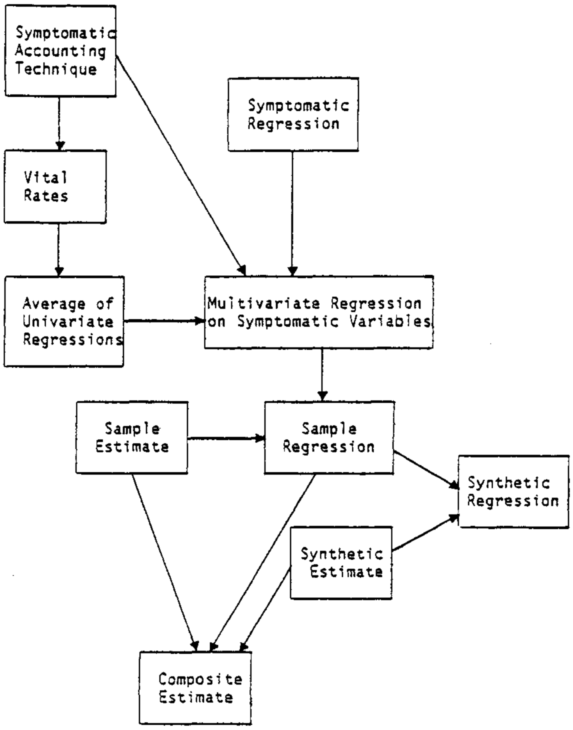
		Information Used		
Techniques	Total Y_{i}.. Estimated By[a/]	Sample Data	Symptomatic Variables X_{ij} (examples)	Assumptions[b/]
Symptomatic Accounting	$\sum_{j=1}^{3} X_{ij} B_j,\ B_j = \pm 1$	No	Births, Deaths, Migration	X_{ij} independent. Regression Coefficients = ± 1.
Vital Rates	$\frac{1}{2}\sum_{j=1}^{2} X_{ijt-1} \frac{X_{sjt}}{X_{sjt-1}}$	No	Births, Deaths	Ratio of local rate to state rate invariant since last census.
Symptomatic Regression	$\sum_{j=1}^{J} X_{ij} B_j$	No	Rates of change of Births, Deaths, etc.	Regression equation invariant since census before the most recent one.
Sample Regression	$\sum_{j=1}^{J} X_{ij} B_j$	Yes (Dependent Variable)	Rates of change of Births, Deaths, etc.	Regression equation since most recent census. Sample data representative of both that area and other areas.
Synthetic Estimation	$\sum_{j=1}^{J} N_{ij} \bar{y}_{.j.}$	Yes (Independent Variable)	Population proportions	Subgroup means are equal in each small area. Population proportions invariant since most recent census.
Synthetic Regression	$\sum_{j=1}^{J} X_{ij} B_j$	Yes (Independent Variable)	Rates of change of Births, Deaths, etc.	Deviations from synthetic estimate are a linear function of the symptomatic variables.

[a/] X_{ij} refers to symptomatic variable j in small area i.

X_{sjt} refers to symptomatic variable j in state s (containing small area i) at time t.

[b/] All techniques make the additional assumptions of multivariate regression.

Figure 1: Interrelationships Among Small Area Estimation Techniques



BIBLIOGRAPHY

Bogue, D.J. "A Technique for Making Extensive Population Estimates." Journal of the American Statistical Association, 45 (June 1950).

_____, and Duncan, B. "A Composite Method for Estimating Postcensal Population of Small Areas by Age, Sex, and Color." Vital Statistics, 47:6 (August 1959).

Ericksen, E. "Recent Developments in Estimation for Local Areas." Proceedings of the Social Statistics Section of the American Statistical Association, 1973.

_____. "A Regression Method for Estimating Population Changes of Local Areas." Journal of the American Statistical Association, 69 (December 1974).

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

Gonzalez, M., and Hoza, C. "Small Area Estimation with Application to Unemployment and Housing Estimates." Journal of the American Statistical Association, 73 (March 1978).

Holt, D.; Smith, T.M.F.: and Tomberlin, T.J. "A Model Based Approach to Estimation for Small Subgroups of a Population." Journal of the American Statistical Association, 74 (June 1979).

Levy, P.S. "The Use of Mortality Data in Evaluating Synthetic Estimates." Proceedings of the Social Statistics Section of the American Statistical Association, 1971.

_____. "Small Area Estimation--Synthetic and Other Procedures, 1968-1978." National Institute on Drug Abuse Research Monograph 24, 1978.

Schaible, W. "A Composite Estimator for Small Area Statistics." National Institute on Drug Abuse Research Monograph 24, 1978a.

_____. "Choosing Weights for Composite Estimators for Small Area Statistics." Proceedings of the Section on Survey Research Methods of the American Statistical Association, 1978b.

_____; Brock, D.; and Schnack, G. "An Empirical Comparison of the Simple Inflation, Synthetic, and Composite Estimators for Small Area Statistics." Proceedings of the Social Statistics Section of the American Statistical Association, 1977.

Schmitt, R., and Crosetti, A. "Accuracy of the Ratio-Correlation Method for Estimating Post-Censal Population." Land Economics, (August 1954).

U.S. Department of Commerce, Bureau of the Census. Estimates of the Population of States with Components of Change, 1970 to 1973. Current Population Reports, Population Estimates and Projections, Series P-25, 520 (July 1974).

_____. Population and Per Capita Money Income Estimates for Local Areas: Detailed Methodology and Evaluation. Current Population Reports, Population Estimates and Projections, Series P-25, 699 (June 1980).