SMALL AREA ESTIMATION PROSPECTS FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Steven G. Heeringa, Survey Research Center, University of Michigan

I. Introduction

The Survey Research Center of the University of Michigan is working with the Income Survey Development Program to examine the use of small area estimation methods with the Survey of Income and Program Participation (SIPP). Kasprzyk and Lininger (1981) provide a description of the proposed design and contents of the 1982 SIPP. This paper discusses the methodology and preliminary results for our first empirical tests of synthetic estimation of state level characteristics based on sample data from the 1979 Research Panel, a national longitudinal survey of the ISDP.

Our current research is examining the application of small domain estimation techniques with 1979 Research Panel data to produce state level estimates of population totals and characteristics for program beneficiaries, classes of individuals such as the poor, elderly or disabled and other populations of current policy interest.

It should be noted within the set of state domains, there is considerable variation in individual sizes; California contains roughly 1/10th of the United States' population, Alaska only 1/565th. The disparity in size becomes even greater when considering estimates for subclasses of states' populations. Such variation in sizes and population characteristics suggests that a single small area estimation method may not be best for each of the individual states. For simplicity, in the preliminary investigations we apply a given method uniformly across all states. Future research will investigate the combined use of methods to produce best results for individual states.

II. Initial Investigations of Small Area Methods

Several estimation methods appear to hold promise for use with the SIPP data: synthetic methods, regression techniques (Ericksen, 1974), the composite estimator (Schaible, 1979) and possibly the empirical Bayes (James-Stein) approach discussed by Fay and Herriot (1979). In our research, the initial focus has been placed on synthetic estimation of small area totals and frequencies for populations of interest to the SIPP program.

As presented in the earlier literature, the synthetic estimator has intuitive appeal and is straightforward to use; however, a criticism of this early treatment is that it lacks a framework for analytical assessment of the synthetic estimator's statistical properties. For the synthetic estimation of small area frequencies and totals, the categorical data approach (Purcell, 1979) provides such a statistical framework. The following paragraphs provide a brief review of the synthetic estimator, introduce the categorical data approach to synthetic estimation and attempt to demonstrate the linkage between the two nominally different methods.

Synthetic Estimators

Synthetic estimates first appeared in the 1968 publication of the National Center for Health Statistics (NCHS), Synthetic State Estimates of Disability, a report describing the new method's application to state level estimates of long- and short-term disability. In subsequent work, NCHS continued its investigation of the synthetic and related estimators, applying and testing the methods for state estimates of mortality, disabilities, and general health characteristics. In the late 1960's, the Bureau of Census began its own investigation into the use of synthetic estimators for calculating small area estimates of dilapidated housing (Gonzalez, 1973) and unemployment rates (Gonzalez and Hoza, 1978).

The application of the synthetic method to small area estimation has not been restricted to the United States. Laake and Langva (1976) - Norway, Purcell and Linacre (1976) - Australia and Ghangurde and Singh (1977) - Canada have applied the synthetic method to data from national household samples.

A general form for the synthetic estimator is:

$$Y_h^* = \sum_g W_{hg} \bar{y}_{.g} \qquad (1.1)$$

where: W_{hg} = an associated variable weight factor for subpopulation g of small area h; and

$\bar{y}_{.g}$ = a consistent large domain or total sample estimate of the characteristic mean or proportion for subpopulation g.

(Note that if $\bar{y}_{hg}$ is substituted for $\bar{y}_{.g}$ in formula (1.1) the result is the conventional post-stratified estimator - a possible explanation for the intuitive appeal of the synthetic estimator). In most applications, the associated variable weight, W_{hg}, is a population total and for the rest of this paper N_{hg} will be used in place of W_{hg}. The sample statistics, $\bar{y}_{.g}$, are usually estimates of subpopulation proportions; e.g. the proportion of persons in the g^{th} subpopulation who are currently employed.

As the following expression for its expected value indicates, the synthetic estimator may be biased.

$$E(Y_h^*) = Y_h + \sum_g N_{hg} (\bar{Y}_{.g} - \bar{Y}_{hg}) \qquad (1.2)$$

The implicit model underlying the synthetic estimator is $\bar{Y}_{hg} = \bar{Y}_{.g}$ for all $g = 1,\ldots,G$ subpopulations. If the model holds, the synthetic estimator will be unbiased for Y_h, the true value of the small area statistic. In practice, there are likely to be departures from the model and the synthetic estimator will have a bias which is a summation of the weighted differences between $\bar{Y}_{.g}$ and $\bar{Y}_{hg}$.

The mean square error of the synthetic estimator can be expressed as:

$$MSE(\bar{Y}^*_h) = \sum_g N^2_{hg} \sigma^2_{\bar{Y}_{.g}} + [\sum_g N_{hg}(\bar{Y}_{.g} - \bar{Y}_{hg})]^2 \quad (1.3)$$

where N_{hg} is the associated population total for the hg cell and $\sigma^2_{\bar{Y}_{.g}}$ is the variance of the subpopulation mean, $\bar{Y}_{.g}$.
The first of the two mean square error terms is the variance component; the second, the square of the estimator bias. For synthetic estimates based on large samples, the variance component -- which is estimable -- will be small relative to that for the direct or post-stratified estimator, but the bias term may be large. Since the bias cannot be reliably measured, the MSE cannot be estimated from the sample data.

The general form of the synthetic estimator appears in the literature as two different expressions.

A first form is a synthetic estimator of small area means, proportions and other ratios.

$$\bar{Y}^*_h = \sum_g N_{hg} \bar{y}_{.g} / N_{h.} \quad (1.4)$$

where $N_{h.}$ is the associated population total for small area h. This is the form of the synthetic estimator which NCHS introduced in 1968. The focus of extensive empirical testing, it is the "synthetic" estimator which appears most frequently in the literature.

The second form, which is an estimator of totals for small area population characteristics, is given by Purcell and Linacre (1976):

$$Y^*_h = \sum_g N_{hg} \bar{y}_{.g} \quad (1.5)$$

or alternatively

$$Y^*_h = \sum_g N_{hg} Y'_{.g} / N_{.g}$$

where $Y'_{.g}$ in the alternative form is the sample estimate of the total of y for subpopulation g. This particular estimator is the form of interest in this paper. Purcell (1979) labels it the BASE estimator (basic synthetic estimator).

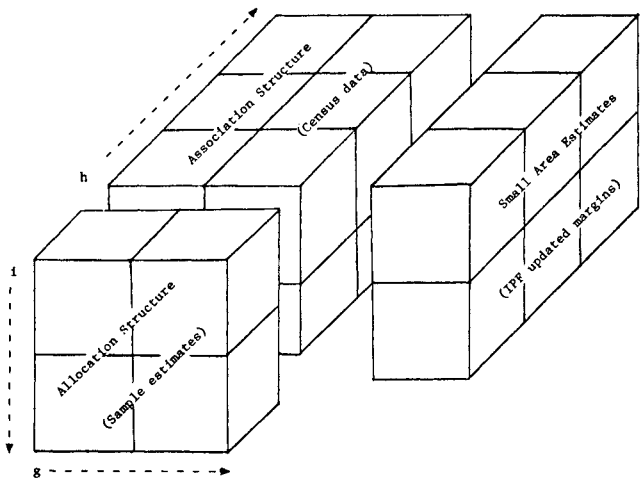
The Categorical Data Approach

Purcell and Kish (1980) propose that synthetic estimation of small area frequencies and totals be addressed in a categorical data framework.

The categorical data approach requires two basic types of information: i) the association structure--data often from a previous census which establishes for small domains the relationship between the variable of interest and a set of associated variables, and ii)the allocation structure--current data (typically from a large survey) which updates this relationship at the level of a larger domain such as the nation or a major Census region.

The association structure is best represented as a three dimensional contingency table with cell frequencies or population counts denoted by $\{N_{hig}\}$ where h = 1,...,H is the small area subscript, i = 1,...,I are the levels of the variable of interest (e.g. employed, unemployed) and g = 1,...,G are subgroups into which the population is classified. (Figure 1 provides a visual representation). The detailed data of the association structure specifies for some past point in time: i) the relative sizes of the N_{hig} and ii) any interactions which existed between small areas (h), subpopulations (g) and levels of the variable of interest (i).

Figure 1: Schematic representation of the categorical data method for small area estimation.



The allocation structure is comprised of a set of one or more updated margins for the association structure. These updated margins may be estimates from sample survey data, statistics from auxiliary data such as administrative records or a combination of these data sources. Shown schematically in Figure 1, the allocation structure updates the ig margin of the association structure; $M = \{m_{.ig}\}$. In practice, this form of the allocation structure will be prevalent; however, there are other possibilities. Examples of other allocation structures are $M = (\{m_{.ig}\}, \{m_{h..}\})$ or even $M = (\{m_{.ig}\}, \{m_{h.g}\})$.

Given this general statistical framework the first step in the estimation process is to update the association structure, making it conform to the new allocation structure. Estimation of cell frequencies for contingency tables with fixed margins was first discussed by Deming and Stephan (1940), who show that estimates (in our case the updated association structure) can be obtained using iterative proportional fitting (IPF), an iterative method which approximates the least squares solution. The form of the derived estimator is dependent on the set of marginal constraints (allocation structure) that is applied.

After iterative proportional fitting has been used to update the association structure, the estimate of the total for level i of small area h is obtained by summing the corresponding hig cells over the g dimension:

$$\hat{Y}_{hi} = \sum_g N'_{hig}, \quad (2.1)$$

where N'_{hig} is the updated estimate of the frequency for the hig cell of the association

structure. $\hat{Y}_{hi}$ is the hi marginal ($m_{hi.}$) of the updated association structure (see Figure 1).

In estimation problems such as those which will be encountered with SIPP data, the detailed data needed for a complete association structure may not be available. Incomplete association structures can be used, but it is necessary to adopt a model of interactions to create a dummy association structure with a full complement of h x i x g cells.

The following example illustrates the use of an incomplete association structure and also brings out the relationship between the categorical data approach and the synthetic estimator of small area totals. Let the association structure be defined by $N = \{N_{hig}\}$, $(h = 1,\ldots,H,\ g = 1,\ldots,G)$; there is no information on the distribution of the characteristic of interest either within the small area or across the subpopulation categories. From current survey estimates, $\{\bar{Y}_{.ig},\ i = 1,\ldots,I,\ g = 1,\ldots,G\}$ an allocation structure, $M = \{m_{.ig}\}$, is developed. A dummy association structure is then defined by assuming a model of proportionality across the i levels of the variable of interest,

$$N_{hig} = N_{h.g}\bar{y}_{.ig}$$

From Deming and Stephan (1940), the least squares estimates for the updated cell frequencies for this case are of the following form:

$$N'_{hig} = N_{hig}\bar{Y}'_{.ig}/N_{.ig} \qquad (2.2)$$

The small area estimator 2.1 then becomes:

$$\hat{Y}_{hi} = \sum_g N_{hig}\bar{Y}'_{.ig}/N_{.ig} \qquad (2.3)$$

Suppressing the i subscript, this expression (2.3) is equivalent to the alternative form of the synthetic estimator given in (1.5).

III. Empirical Investigation

The empirical investigation of small area methods reported here uses data from the 1979 ISDP Research Panel in estimating state subpopulation totals for:

1. Social Security beneficiaries,
   a. retired workers,
   b. disabled persons;
2. Basic Education Opportunity Grant (BEOG) awardees; and
3. Union members.

For each of these test estimates, current population values are available from an exogenous source. 1979 state totals for social security recipients are available in the administrative records summaries published in the Annual Statistical Supplement to the Social Security Bulletin. BEOG awards summaries by state for 1979 were obtained via a request to the U.S. Department of Education. State's union membership was extracted from the 1979 Statistical Abstract of the United States.

The auxiliary data used in synthetic estimation exert a critical influence on the accuracy of the small area estimates. For estimating 1979 state totals of Social Security recipients and BEOG awardees, state population totals for categories of age, race, and sex form the association structure. States' employment for major industry and occupation categories comprises the association structure for the estimation of states' union membership. The 1970 Census is the source of the age by race by sex association structure for the results reported in this paper. Occupation and industry data are available from the 1970 Census public use sample; however, in our investigations we chose to go directly to individual state government agencies to obtain the most current data on employment by occupation and industry.

Using the categorical data method, two related "synthetic" estimators are being evaluated: i) the BASE estimator (see 1.5), and ii) the BASE estimator with a control to 1980 total population for each state. The state population control adds information to the allocation structure in the form of the vector of marginal constraints $\{m_{h..}\}$. The association and allocation structures for the two estimators are given in the following table.

	Estimator	Association Structure	Allocation Structure
I)	BASE (1.5)	$N=\{N_{h.g}\}$	$M=(\{m_{.ig}\})$
II)	BASE plus 1980 state population control	$N=\{N_{h.g}\}$	$M=(\{m_{.ig}\},\{m_{h..}\})$

To solve for state estimates, the iterative proportional fitting program documented in Purcell (1979) is being used. Only slight modifications were made to the original FORTRAN code to facilitate data input and to adapt the output to our research needs.

To evaluate the performance of the estimators, we compared the estimated statistics for states to the population values obtained from exogenous sources. For simplicity of interpretation, we focus on absolute relative error ($ARE = |\hat{Y}_h - Y_h|/Y_h$) of estimates in this report. In this regard an important point should be made. ARE's reflect the magnitude of the error only in relative terms and summary statistics (median ARE) take no account of the sizes of individual states.

IV. Preliminary Findings.

Several observations on the results of the empirical tests of the synthetic estimators are immediate: i) the overall level of accuracy for a set of state estimates varies considerably for the selected test items (see Table 1), ii) for a test item, the level of accuracy for individual states also is highly variable, and iii) the addition of population controls -- the h margin of the allocation structure -- results in improved accuracy except for estimates of state's union membership.

Table 1: Synthetic Estimation of Totals for Selected Subpopulations of the 50 states and the District of Columbia. Descriptive statistics for estimates' absolute, relative error: ARE = $|\hat{Y}_h - Y_h|/Y_h$

Subpopulation	Absolute Relative Error		
	Median	Strd. Dev.	Maximum
Soc. Sec.: Ret.	.074	.08	.352:Nevada
Soc. Sec.: Dis.	.184	.18	.753:D.of Col.
BEOG Awards	.148	.17	.633:Indiana
Union Members	.329	.51	2.177:N.Car.
With 1980 Population Control			
Soc. Sec.: Ret.	.051	.06	.203:Louisiana
Soc. Sec.: Dis.	.145	.15	.543:Minnesota
BEOG Awards	.128	.21	1.009:Nevada
Union Members	.343	.60	3.302:N.Car.

As seen in Table 1, absolute relative errors for estimates of retired social security beneficiaries are lower on the average than those for the three other test items, reflecting in part the strong relationship between the age variable of the association structure and eligibility for retirement benefits. Since eligibility is almost totally dependent on a worker's age and program coverage is nearly universal, the model underlying the synthetic estimator should be expected to hold reasonably well. Median ARE for state estimates of social security retirement beneficiaries is .074 for the BASE estimator, .051 when 1980 state population constraints are added to the allocation structure. Focusing on individual states, BASE estimates for Arizona, Florida and Nevada are considerably less than the actual values of numbers of social security retirees; however, introduction of a 1980 population control to the estimator reduces these extreme errors to levels comparable to those for other states.

Estimates for total numbers of disabled social security beneficiaries by state are less accurate overall than those for retired beneficiaries. The median ARE is .184 for BASE estimates of states' disabled beneficiaries and .145 for population controlled values. ARE's for individual states show no noticeable pattern except that the introduction of the 1980 population control typically reduces the ARE of the estimate of disabled beneficiaries.

The general level of accuracy for estimates of Basic Educational Opportunity Grant awards made to state's residents is comparable to that for social security disability benefits. For the BEOG test item, the association structure variables, age and race, are linked to BEOG eligibility, awards going primarily to college age persons and due to income criteria, disproportionately to students who are black. The accuracy of the synthetic estimator is probably influenced by differences in college attendance rates among the states.

Union membership for states is the test item on which there appears to be a clear breakdown of the model underlying this application of the synthetic method. The assumption which is made is that the proportion of unionized members in an occupation/industry class is identical for states and the nation at large. Using an allocation structure--sample estimates--based on national sample data, union membership for Southern and other traditionally non-union states is grossly overestimated. The addition of a control to current labor force estimates only exaggerates the errors for these states since they are among the highest employment growth areas of the nation. Median ARE's reflect the poor performance of the synthetic estimates for this test item; BASE (.329), BASE + control (.343). The largest ARE's for states' BASE estimates of union membership were as noted in the South--North Carolina (2.18), South Carolina (2.17), Florida (1.35), Texas (1.32), Georgia (.90). Substantial errors are also observed for the plains and mountain states.

For individual states, the level of accuracy of the synthetic estimates does not show a strong consistency across the four test items. Notable exceptions are the states of California and New Jersey for which estimation errors were generally very low. On the three recipiency items, estimation errors for rural New England states, the Dakotas, the mountain states, Alaska and Hawaii were high relative to those for other states. As noted previously, estimates of union membership showed large errors in the South and rural midwest and mountain states.

Table 2 presents individual results for the nation's ten largest states. With the individual exceptions noted above, the collective results for the nation's ten largest states are similar to those for the combined set of all states. Since the accuracy of the synthetic method is a function of the degree to which the implicit model, $\bar{Y}_{.g} = \bar{Y}_{hg}$, holds and not the population size estimation unit, this finding is not surprising.

The errors observed in the empirical estimates are a function of i) the estimator variance which enters through the sampling error of the allocation structure and ii) the bias which results from a breakdown of the model underlying the estimator. Gonzalez and Hoza (1978) and others note that synthetic estimates exhibit a shrinkage from the true value toward a general mean for the population ($\bar{Y}..$), errors for states in which the subpopulation or characteristic of interest is less prevalent than average ($\bar{Y}_{h.} < \bar{Y}..$) will tend to be positive (overestimates). Negative errors (underestimates) are typically found for states in which $\bar{Y}_{h.} > \bar{Y}..$. This "regression effect" can also be seen in the results of our investigation of the synthetic method, particularly on the test items for which the estimators' performance is poorest.

In our analysis, we observed no significant relationship between a state's population size and the direction or magnitude of the relative error of the synthetic estimates for the set of test items. However, for estimates of social security disability benefits, BEOG awards and union membership, there is a strong relationship between the proportion of a state's population

Table 2: Absolute Relative error for selected test items, BASE AND BASE - controlled estimates for the nation's ten most populous states.

STATE	BASE				BASE + CONTROL			
	RET	DIS	BEOG	UNION	RET	DIS	BEOG	UNION
California	.04	.05	.02	.01	.03	.02	.05	.03
Florida	.25	.32	.15	1.35	.05	.14	.09	1.01
Illinois	.10	.43	.12	.27	.01	.31	.03	.27
Massachusetts	.04	.24	.19	.07	.05	.14	.23	.12
Michigan	.03	.10	.21	.39	.05	.01	.12	.30
New Jersey	.02	.08	.05	.03	.10	.00	.03	.07
New York	.07	.09	.35	.38	.08	.05	.44	.44
Ohio	.12	.11	.45	.27	.03	.01	.33	.20
Pennsylvania	.04	.02	.15	.35	.06	.08	.04	.33
Texas	.04	.18	.07	1.32	.19	.35	.22	1.17
Average ARE	.08	.16	.18	.44	.07	.11	.16	.39

belonging to the subpopulation or characteristic group of interest, $\bar{Y}_{h.}$, and the relative error of the state estimates. The direction of the error is positive (overestimate) when $\bar{Y}_{h.} < \bar{Y}_{..}$ and negative when $\bar{Y}_{h.} > \bar{Y}_{..}$; magnitude of the relative error increases with the distance of $\bar{Y}_{h.}$ from $\bar{Y}_{..}$. For BASE - controlled estimates of states' social security disability beneficiaries, Figure 2 illustrates the observed relationship between absolute relative error (ARE) and $\bar{Y}_{h}$. Absolute relative errors (See Figure 3) for the more accurate estimates of states' social security retirees do not appear to exhibit a regression effect.

FIGURE 2- ABSOLUTE RELATIVE ERROR FOR 1979 STATE ESTIMATES OF SOCIAL SECURITY DISABILITY

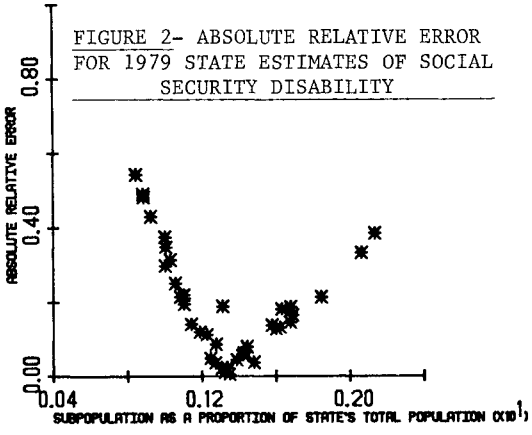
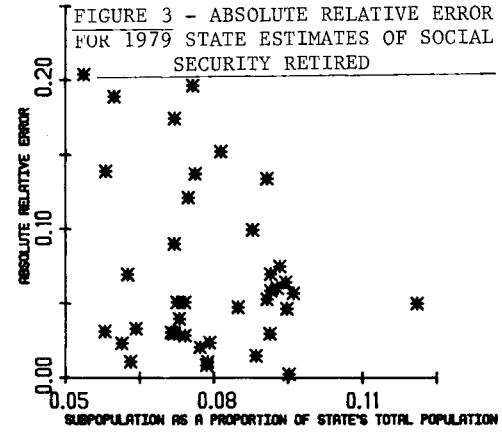


FIGURE 3 - ABSOLUTE RELATIVE ERROR FOR 1979 STATE ESTIMATES OF SOCIAL SECURITY RETIRED



V. Conclusions and Future Work

The limited set of empirical results discussed in the preceding section reinforces the importance of the model in synthetic estimation. In estimating social security retirees, the results seems to indicate that the underlying model holds reasonably well. Since the other three test items experience greater overall errors of estimation, it is likely that the underlying models do not hold as well. For certain combinations of states and test items it is easy to see a reason for the breakdown of the estimator model. In others, the source of a large error is not clear.

Detecting departures from the synthetic estimator model is difficult if not impossible to do at the level of each individual state. In future work we will be investigating methods for determining the approximate degree of model fit by analyzing combined sample data for groups of similar states. This investigation should also provide guidance in determining the best level at which to introduce regional or other geographic controls in the calculation of synthetic estimates.

To varying degrees, the error of the synthetic estimates exhibit a regression effect, suggesting that regression techniques may be effective in reducing the errors of the state level estimates. Several regression methods for small area estimation are described in the literature. The regression-adjusted synthetic estimator described by Levy (1971) is a univariate method for introducing "local" auxiliary data into the estimate. A more complex and possibly more effective option is to introduce several auxiliary variables and the synthetic estimates for a small area as independent variables in a linear multivariate regression estimator (Ericksen, 1974). In subsequent work, we plan to investigate the use of these regression approaches.

For most states, the SIPP sample size will preclude direct estimation of state level characteristics, but in larger states such as California and New York acceptably precise direct estimates may be available. Not wanting to ignore the potential contribution of direct estimation for the largest states, investigation of composite estimators (Schaible, 1979) is also planned.

REFERENCES

Deming, W.E. (1943). Statistical Adjustment of Data. Wiley, New York.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics II. 427-444.

Ericksen, E.P. (1974). A regression method for estimating population changes of local areas. Journal of the American Statistical Association 69, 867-875.

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.

Ghangurde, P.D. and Singh, M.P. (1977). Synthetic estimation in periodic household surveys. Journal of Survey Methodology, Statistics Canada 3, 152-181.

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

Gonzalez, M.E. and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. Journal of the American Statistical Association 73, 7-15.

Heeringa, S. (1981). Small area estimation: prospects for the Survey of Income and Program Participation. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Kasprzyk, D. and Lininger, C. (1981). The 1982 Survey of Income and Program Participation. To be published in the 1981 Proceedings of the Section on Survey Methods, American Statistical Association.

Laake, P. and Langva, H.K. (1976). The estimation of employment within geographical regions: On the bias, variance and mean square error of the estimates. (In Norwegian with English Summary.) Central Bureau of Statistics of Norway, Article 88.

Levy, P. (1971). The use of mortality data in evaluating synthetic estimates. 1971 Proceedings of the Social Statistics Section, American Statistical Association, 328-331.

National Center for Health Statistics (1968). Synthetic State Estimates of Disability. P.H.S. Publication No. 1759. U.S. Government Printing Office, Washington, D.C.

Purcell, N.J. and Linacre, A. (1976). Techniques for the estimation of small area characteristics. Paper Presented at the 3rd Australian Statistical Conference, Melbourne, Australia, 18-20 August, 1976.

Purcell, N.J. (1979). Efficient small domain estimation: A categorical data analysis approach. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

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

Schaible, W.L. (1979). A composite estimator for small area statistics. In Synthetic Estimates for Small Areas (J. Steinberg, ed.), pp. 36-53. National Institute on Drug Abuse Research Monograph 24. U.S. Government Printing Office, Washington, D.C.

Social Security Administration (1980). Social Security Bulletin, Annual Statistical Supplement, 1977-79, U.S. Dept. of Health and Human Services, Washington, D.C.

U.S. Bureau of the Census (1979). Statistical Abstract of the United States:1979. U.S. Government Printing Office, Washington, D.C.