BIOMETRICS 35, 365–384
June, 1979

A BIOMETRICS INVITED PAPER

Estimation for Small Domains

NOEL J. PURCELL[1]

Australian Bureau of Statistics, P.O. Box 10, Belconnen,
A.C.T. 2616, Australia

LESLIE KISH

Survey Research Center, Institute for Social Research,
University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

Summary

Timely and complete health, social and economic data can be obtained from samples, but usually only for major geographic areas and large subgroups of the population. Small domain estimates are available from censuses, but only infrequently and then only for a few variables. Effective planning of health services and other governmental activities cannot depend on traditional data sources; the data must be more current and more complete than these sources provide.

Since estimates are needed for a great diversity of domains, a definition and classification of domains is presented to clarify the direction of this review. The existing small domain estimation techniques are split into several distinct approaches and reviewed separately. The basic methodologies of these techniques are presented together with their data requirements and limitations. The existing techniques are briefly assessed in regard to their performance and their potential for further application. Current research approaches are also reviewed and possible lines for future advances are indicated.

List of Contents

[1] Present address (through August 1979): Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

Key Words: Local area estimation; Synthetic estimates; Small area estimates; Small domain estimates; Postcensal estimates.

1. *Introduction*

Estimates for local areas and other small domains have been of general interest for a long time, but have been unavailable except for estimates from population censuses, special surveys, or administrative registers. These interests, however, have been superseded by increasing demands for more diverse, rich and current data for small domains, which are required for the planning of reforms, welfare and administration in many fields, including health programs. For example, the Planning Act which mandates the Health Systems Agencies (HSA's) specifically requires the HSA's to collect and analyze data relating to the health status of the residents and to the health care delivery systems in their health service areas. Data are also required on such factors as the effects of the delivery systems on the health of the residents, and the environmental and occupational exposure conditions affecting immediate and long-term health conditions. Very recently the demands for data to be used directly for apportioning money and resources have been added (at least in the U.S.A.) to the needs of planners and the curiosity of scientists. These demands are now greatly increasing research interests and efforts in this area.

Estimates for small domains have been largely neglected, until recently, by statistical and sampling theory. As exceptions we note that, for population counts of local areas, statistical demographers have developed several competing methods. But these methods have been essentially accounting procedures specialized for population counts, with the notable addition of the ratio-correlation method, which has wider application (see Section 4). Generally, statistical theory has been concerned with the estimation of overall means based on the entire sample. At the other end, some statistics exist for predictions (and related decision functions) for individual cases. However the problem between the two extremes, estimates for subpopulations (especially small domains), has been largely neglected.

Only in recent years have estimates for small domains become an active area for research. This has resulted in investigations of a variety of statistical techniques for application to problems of estimation for small domains. Apart from regression based procedures, the use of empirical Bayes and of Bayesian methods, of superpopulation prediction theory, of clustering techniques, and of categorical data analysis methods, are being researched. The categorical data analysis approach is of special interest to the authors, as it offers both a structured and logical approach to the problem, through which the properties of the resulting estimators are readily apparent.

Small domain estimates are required for a diversity of domains (subpopulations) and the type of domain may influence the choice of methods. A classification of types of domains is therefore desirable:

(a) Some are *planned domains* for which separate samples have been planned, designed and selected; their combination forms the entire sample; e.g., major regions and other separate strata composed of entire primary units.

(b) At the other extreme are *crossclasses* which cut across the sample design and the sampling units; e.g., age, sex, occupation and education classes widely spread across the sampling units.

(c) Between the two extremes, but less commonly used, are divisions that have not been distinguished in the sample selection but tend to concentrate unevenly in primary units.

Here we follow custom in using *domains* for any subclass or sub-division of populations. It might be better, however, to restrict it in the future to planned domains, as in the United Nations (1950) definition: "Any sub-division about which the enquiry is planned to supply numerical information of known precision may be termed a domain of study."

The sizes of domains also influences the choice of methods, hence we propose a cross-classification of the above types with classes based on the size of the domains, in order to clarify the direction of the later discussion. This subclassification is stated very roughly here to orders of magnitudes, with our descriptive names, and illustrated with common examples.

(1) *Major domains,* comprising perhaps 1/10 of the population or more. Examples: major regions, 10-year age groups, or major categorical classes, like occupations.

(2) *Minor domains,* comprising between 1/10 and 1/100 of the population. Examples: state populations, single years of age, two-fold classifications like occupation by education, or a single small classification like work disability.

(3) *Mini domains,* comprising from 1/100 to 1/10,000 of the population. Examples: populations of counties (more than 3,000 of them in the U.S.A.), or a three-fold classification like age by occupation by education.

(4) *Rare types of individuals,* comprising less than 1/10,000 in the population. Examples: populations of health service areas classified by various ethnic groups, or individuals with specific chronic health problems classified by local area of residence.

The boundaries of these classes should not be taken seriously; they depend on sizes of samples and of populations, on the variables and the statistics, on the precisions and the decisions involved, etc. But they are useful to remind us of practical differences between these types of subclasses and to avoid the common mistake of considering "statistics for states and local areas" as one homogeneous problem.

Probability methods of survey sampling have produced standard estimates, basically without bias, for major domains; however, seldom for minor domains. For planned domains of major size (class 1a), such as major regions, separate and independent samples are commonly designed. For major crossclasses (class 1b), such as 10-year age classes, the proper expectations are provided by probability selections. However, sometimes for minor domains (class 2) and usually for mini domains (class 3), the standard methods of survey estimation break down, because the sample bases are ordinarily too small for any usable reliability, and new methods are needed. For truly rare items (class 4), sample surveys are usually useless; separate and distinct methods are required. Thus small domain methods reviewed here are directed principally to classes 2 and 3, and rarely to class 4, and the distinction may help to cover the wide range implied. Although demand for "small area statistics" (classes 2a and 3a) are most prominent currently, small crossclasses (classes 2b and 3b) are also important.

It is most useful to classify the existing methods, as reviewed in Sections 2 through 7, by the sources of data on which they rely, and Table 1 does that. Each of the main sources—censuses, registers, samples—lacks at least one of the key requirements: detail, timeliness, and relevance. Census data are detailed, but not timely, and sometimes not relevant either. Symptomatic data from registers can be timely, and detailed, but usually not entirely relevant. Sample data can be made to fit our needs, but cannot provide the needed detail.

There are also approaches that use only sample data, such as the "double sampling-regression" approach reported by Hansen, Hurwitz and Madow (1953), which was in a sense a forerunner of the sample-regression method discussed in Section 5. Bayes and empirical Bayes approaches can also be used when only sample data are available, and reference is made to these methods in Sub-section 9.3.

In Section 8, existing methods are briefly appraised. Current research and the lines of

TABLE 1
Existing Methods Classified by Sources of Data

Section	Methods	Census	Register	Sample
2	Symptomatic Accounting	*	*	
3	Synthetic (Ratio)	*		*
4	Regression-Symptomatic	*	*	
5	Sample-Regression	*	*	*
6	Synthetic-Regression	*	*	*
7	Base Unit	*	*	*

future advances are then discussed in Section 9, with several promising approaches being reviewed, which are the subject of considerable research interest. Finally, in Section 10, we make some summary remarks.

Three related matters deserve the briefest mention. First, that when repeated surveys are available, their data can be accumulated for more precision for small domains; e.g., monthly surveys can be accumulated to give more accurate yearly data. Second, to obtain and properly utilize information from administrative registers can be the key to good estimation at the small domain level; note the diverse methods of Section 2. Third, there exist other methods for improving the search for "rare items" (class 4); see Kish (1965), Section 11.4.

2. *Symptomatic Accounting Techniques*

The standard symptomatic accounting techniques (SAT) are the oldest of the small domain estimation techniques and essentially use logical demographic relationships in combination with statistical relationships based on previous data. This is the current approach followed for local population estimation in the United States, as detailed by the U.S. Bureau of the Census (1975*a*). Basic demographic accounting equations relate births, deaths, and migration to change in population. In addition, other equations are used in these procedures which relate the growth in population to growth in symptomatic variables, such as the numbers of births and of deaths, of dwellings, of school enrollments, of income tax returns, etc. As a general rule, these latter relationships are developed and validated by the use of census data.

The SAT techniques, widely referred to as "methods," have two limitations. First, they depend on good current registers of births, deaths, etc.; second they are used only for population counts with which the symptomatic variables are strongly correlated. Since our chief interest lies in more portable techniques, we restrict ourselves to a brief mention of the essential features of the techniques and provide references to detailed accounts of the particular methodologies. Good reviews of most of these methods can be found in Ericksen (1971), Kalsbeek (1973), and U.S. Bureau of the Census (1975*b*).

The estimation of the net civilian migration component is the primary objective of the Census Bureau's *component methods,* which were originally proposed by Eldridge as stated in U.S. Bureau of the Census (1947). Additionally, the methods take direct account of natural increases and the net loss to the armed forces. There are two alternate component methods, both of which base their estimate of net civilian migration in the postcensal period on the assumption that the migration rate for the total population is the same as the migration rate for school children. The methods differ in the way the migration rate of school children is estimated. A detailed description of the *census component method I* is given by the U.S.

Bureau of the Census (1949) and of the *census component method II* by the U.S. Bureau of the Census (1966).

A somewhat different procedure for estimating postcensal population sizes is given by the *vital rates technique,* of which the first full description was given by Bogue (1950). This approach uses large area birth and death rates and a censal-ratio procedure to estimate the required local area population sizes. The basic assumption underlying the vital rates technique is that the local area birth and death rates have changed by the same proportion as the large area rates.

The *composite method,* whose development is attributed to Bogue and Duncan (1959), was devised as an alternative to the vital rates technique. The method divides the local area population into distinct age groups and obtains population estimates for each of these groups separately, using the techniques and data considered most appropriate for estimating each of these subgroups. The resulting subgroup estimates are then summed to give a final local population estimate.

A method that is often used to make population estimates for cities and metropolitan counties is the *housing unit method.* This method uses current estimates of the number of housing units in the local areas and the average number of individuals per housing unit, and is based on the assumption that changes in the number of housing units reflect changes in populations. The U.S. Bureau of the Census (1969) discusses this method and reports on a modification of this method to resemble a composite method. More recently Rives (1976) presented a further modification of the housing unit method.

The newest of the standard symptomatic accounting techniques is the *administrative data records method,* which has been introduced by the U.S. Bureau of the Census, principally for updating population estimates for all revenue sharing areas as required for the Federal Revenue Sharing program. Basically, the method is similar to the census component methods, except that the estimate of net migration is based on numbers of incomes filed with the Internal Revenue Service (IRS). The use of individual records, rather than classes of data, permits estimates for very small areas; however it is limited to Federal use, due to confidentiality provisions. A detailed description of the method, indicating its strengths and pointing out some of the problems in implementing such a procedure, is given by Starsinic (1974).

3. Synthetic Estimation

Synthetic estimation uses sample data to estimate, at some higher level of aggregation, the variable of interest for different subclasses of the population; then it scales these estimates in proportion to the subclass incidence within the small domains of interest. For example, state estimates of unemployment, cross-tabulated by age, sex and race, might be scaled by the proportional incidence of these subclasses in each county to estimate county unemployment. The estimates will be correct if the composition of each county is known accurately (perhaps from census data) and if the state unemployment rates, for each demographic subgroup, correctly reflect the subgroup unemployment rates in each county.

This approach was given the name synthetic estimation by the National Center for Health Statistics (1968), in what seems to be the first documented use. In this case the method was used to calculate state estimates of long and short term physical disabilities, based on the National Health Interview Survey. The term *synthetic* was used because these estimates were not derived directly from survey results. However, this term is now used, more specifically, to refer to this particular method of borrowing information from similar small domains in order to increase the accuracy of the resulting estimates.

More recently synthetic estimation has been used in connection with a number of surveys to obtain estimates of small domain characteristics. Gonzalez (1973) reports on work carried out at the U.S. Bureau of the Census where synthetic estimates were used to revise the population count of the 1970 Census of Population and Housing for the population of housing units reported as vacant but actually occupied. In a paper that concentrates on measuring the errors of synthetic estimates, Gonzalez and Waksberg (1973) discuss the production of county unemployment rates from data for regional estimates obtained from the Current Population Survey. Gonzalez and Hoza (1978) have made a more extensive study of the use of synthetic methods to estimate unemployment rates for counties, as well as the production of estimates of dilapidated housing units. Synthetic estimates of unemployment have also been studied by Schaible, Brock and Schnack (1977), who compared the average squared errors of synthetic and direct estimates of unemployment rates for county groups in Texas. Purcell and Linacre (1976) discussed two empirical studies which were carried out at the Australian Bureau of Statistics, aimed at the production of synthetic estimates of income and work force status for Australian Census Statistical Divisions. Considerable work has also been carried out at Statistics Canada towards the regular production of synthetic small domain estimates from the Canadian Labour Force Survey (Ghangurde and Singh 1978), and at the Central Bureau of Statistics of Norway towards the production of synthetic estimates of employment (Laake and Langva 1976, Laake 1978).

Synthetic estimation has additionally been important in the public health field. Levy (1971) reported on an evaluation of synthetic state estimates of the number of deaths from four different causes. These estimates were evaluated by comparison with the official state mortality statistics. Namekata (1974) and Namekata, Levy and O'Rourke (1975), using data from the 1970 census, have studied synthetic estimates of complete and partial work disabilities for states. More recently, the National Center for Health Statistics (1977*a*) published a new set of synthetic state estimates of disability, and utilization of medical services. Finally, a recent report of the National Center for Health Statistics (1977*b*) examined synthetic methods for estimating health characteristics for individual states.

The synthetic method can be formalized as follows. Suppose we wish to estimate a characteristic x within a number of small domains. Estimates are assumed to be available from survey data for the characteristic x, cross-classified by non-overlapping and exhaustive subgroups of the population only for some larger domain that encompasses the smaller domains. From some past data source, usually a previous census, we also have information on some associated variable(s), Y, classified by the same non-overlapping and exhaustive subgroups of the population as above. That is, we have available the following information: the count for the associated variable for the *hth* small domain and *gth* subgroup, Y_{hg}, and the survey estimate for the characteristic x for subgroup g at the large domain level, based on the sample, hence the prime, $x'_{.g}$.

Our synthetic estimate of the total for characteristic x in small domain h is then

$$\hat{x}_h = \sum_g \hat{x}_{hg} = \sum_g (Y_{hg}/Y_{.g})x'_{.g}, \tag{3.1}$$

where the dot subscript is used to denote summation over that subscript. Usually the associated variable Y is taken to be the number of people so $Y_{hg} = N_{hg}$ and our synthetic estimate becomes $\hat{x}_h = \sum_g(N_{hg}/N_{.g})x'_{.g}$. The synthetic estimator (3.1) may be viewed as an extension of the basic ratio estimator to g groups. Additionally, it has the desirable property that it corresponds to the large domain estimate when summed over exhaustive small domains:

$$\sum_h \hat{x}_h = \sum_h \sum_g (Y_{hg}/Y_{.g})x'_{.g} = \sum_g \sum_h (Y_{hg}/Y_{.g})x'_{.g} = \sum_g x'_{.g} = x'_{..}$$

The estimator (3.1) proposed by Purcell and Linacre (1976) and later by Ghangurde and Singh (1977) is not the estimator proposed by the National Center for Health Statistics (1968), and investigated by Gonzalez (1973) and others. That estimator is of the form

$$\bar{x}_h = \sum_g (Y_{hg}/Y_{h.})\bar{x}'_{.g}, \tag{3.2}$$

where $\bar{x}_h$ is the synthetic estimator of the small domain mean for characteristic x, and $\bar{x}'_{.g}$ is the survey estimate of the mean for the characteristic x for subgroup g at the large domain level. In general (3.2) does not proportionally add to the corresponding unbiased large area estimate as the estimator (3.1) does. Both estimators use Y_{hg}, the size of the associated variable count for the small domains; but (3.2) uses it for the ratio of the subgroups within the small domains, whereas (3.1) uses it for the ratio of the small domains within the subgroups.

Synthetic estimates reduce variances, but they are biased estimates for two reasons. First, there will often exist departures from the underlying assumption of homogenity of rates. Second, the weights $Y_{hg}/Y_{.g}$ are usually based on past data, and the structure of the population may have changed during this time. Looking at the bias expression for the estimator (3.1) we have, where X denotes the true value of the characteristic x,

$$E[\hat{x}_h - X_h] = E\left[\sum_g (Y_{hg}/Y_{.g})x'_{.g} - \sum_g X_{hg}\right] = \sum_g Y_{hg}(X_{.g}/Y_{.g} - X_{hg}/Y_{hg}).$$

So the synthetic estimator is biased for X_h unless $X_{.g}/Y_{.g} = X_{hg}/Y_{hg}$, for all subgroups. This will not hold in general.

Evaluating the synthetic estimates is complicated by the bias, hence attention is directed to the mean square error (MSE) of the estimates, but this is difficult to estimate, due to the lack of knowledge about the true X_{hg} values needed to estimate the bias term. This problem can theoretically be overcome in situations where it is possible to form some unbiased sample estimates at the small domain level, x'_{hg}, even if they have large sampling variability. However, the small domain variance estimates are likely to be unstable. As a possible solution to this problem, Gonzalez and Waksberg (1973) suggested the use of the average mean square error over the small domains of interest. Their proposed *average MSE* is given by $\sum_h E(\hat{x}_h - X_h)^2/H$, where H is the number of small domains, and an estimate of this average MSE is derived under some limiting assumptions.

A different approach, that does not depend on having an unbiased estimate of x'_{hg} available, is to use census data for the evaluation, assuming the variable of interest x has been collected in the census. Ghangurde and Singh (1978), for example, in considering the problem of evaluating the efficiency of synthetic estimates based on cluster sampling with probability proportional to size selection, used census data to estimate the parameters in the resulting bias and variance expressions. These expressions are derived in a framework of super-population models.

One of the difficulties with the synthetic estimates is that, unless the grouping variables are highly correlated with the variable of interest, the synthetic estimates will tend to cluster near the mean for the large domain, and fail to reflect the actual effects of local area factors. For small domains with divergent values the synthetic estimates can be poor. However, with careful choice of grouping variables the synthetic estimator has been demonstrated to lead to usable results (see, for example, Gonzalez 1973, Purcell and Linacre 1976, Namekata 1974). The main advantage of the method being its ease of calculation.

The term synthetic estimation has also been used loosely to apply to some early work on small domain estimation in the public opinion research field. This work lacks the rigor of other techniques reviewed here and consequently will not be discussed. A review of this approach and extensions of it can be found in Cohen (1978).

4. *Regression-Symptomatic Procedures*

The distinguishing feature of the regression-symptomatic methods is that they are based on the fitting of a functional relationship (least squares regression) between the variable of interest and the symptomatic variables. The small area estimates are then obtained from this fitted model using current information.

Multiple regression has become an important tool for making small domain estimates since the reintroduction and modification of the *ratio-correlation method* by Schmitt and Crosetti (1954). The ratio-correlation technique was first proposed by Snow (1911). Basic to the method is the assumption that the same relationships between the symptomatic indicators and the variable of interest, computed for the intercensal period, also hold in the postcensal period. Briefly, the method is as follows:

(a) For each of p symptomatic variables, Y_i, the proportion of the total which belongs to each small domain, h, is calculated for times $t = 1, 2$, the two base periods (usually census dates), and for 3, the forecast data. We shall denote these proportions as $P_{hti} = Y_{hti}/\sum_h Y_{hti}$, where Y_{hti} denotes the value for the *ith* symptomatic variable in small domain h and at time t. Examples of symptomatic variables that have been used include births, deaths, employment, voter registration, school enrollments and so on.

(b) Similarly for the variable of interest, x, we calculate the proportions in each small domain, for times 1 and 2, $q_{ht} = x_{ht}/\sum_h x_{ht}$.

(c) Two sets of ratios are then calculated. First, reflecting the changes between times 1 and 2, we compute

$$r_{hai} = P_{h2i}/P_{h1i}, \qquad R_{ha} = q_{h2}/q_{h1}, \qquad i = 1, 2, \ldots, p. \tag{4.1}$$

The second set of ratios is constructed to show the change between times 2 and 3 in the symptomatic variables. That is

$$r_{hbi} = P_{h3i}/P_{h2i}, \qquad i = 1, 2, \ldots, p. \tag{4.2}$$

(d) Using multiple regression, the functional model is fitted with the ratio variables given in (4.1). That is we fit

$$R_{ha} = \hat{B}_0 + \hat{B}_1 r_{ha1} + \ldots + \hat{B}_p r_{hap},$$

where $\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_p$ are the regression coefficients, estimated from changes between the two base periods.

(e) The coefficients $\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_p$, which are based on the empirical relationships observed in the intercensal ratios, are then used with the postcensal ratios, given in (4.2) for the symptomatic variables, to predict the ratio of population proportions for the postcensal period. The predicted ratios are therefore given by

$$\hat{R}_{hb} = \hat{B}_0 + \hat{B}_1 r_{hb1} + \ldots + \hat{B}_p r_{hbp}.$$

(f) The predicted ratios $\hat{R}_{hb}$ can then be translated into actual numbers by multiplying these ratios with the actual small domain proportions, for the variable of interest at the previous census, and with its current value summed over all small domains. Thus,

$$\hat{x}_{h3} = \hat{R}_{hb}\left(x_{h2}\Big/\sum_h x_{h2}\right)\sum_h x_{h3}.$$

This method depends on the assumption that the observed statistical relationship between the independent and dependent variables in the last intercensal period will persist in the current postcensal period. The adequacy of this assumption is dependent on the size of the multiple correlation, hence on the number and combined value of symptomatic variables used, as well as on the stability of the relationships over time. In most cases, considerably more symptomatic information is required than for other methods. Namboodiri (1972) presents a number of other theoretical problems associated with the ratio-correlation method; he also demonstrates a procedure for the averaging of the results of univariate regression estimates, which improves the accuracy of the ratio-correlation method. Another extension of the ratio-correlation approach has been presented by Rosenberg (1968), based on estimates of different relationships within separate strata. Pursell (1970) used a similar procedure, but also added dummy variables to the set of symptomatic indicators. Both studies also found improved intercensal estimates; however, postcensal estimates were not computed.

The ratio-correlation method has been used chiefly for estimates of total population, but it can be used for other statistics. Martin and Serow (1978) have applied the ratio-correlation method to the estimation of the age and race compositions of populations at sub-state levels. In addition they studied the effects of the extensions to the basic procedure mentioned above. In the case of local areas in Virginia which they studied, they surprisingly found that the extensions failed to produce estimates that were clearly and consistently superior to the basic ratio-correlation procedure.

A variation of the ratio-correlation method has been presented by O'Hare (1976), which is best termed the *difference-correlation method.* Its distinction is in the construction of the variables which are used to reflect change over time. Differences between the proportions at the two pairs of time points given in (4.1) and (4.2) are used rather than their ratios. That is, this method uses

$$d_{hai} = P_{h2i} - P_{h1i}, \quad d_{hbi} = P_{h3i} - P_{h2i}, \quad \textit{and} \quad D_{ha} = q_{h2} - q_{h1}$$

in place of r_{hai}, r_{hbi} and R_{ha} in the multiple regression equation. O'Hare claims for the data he used, that the structure of relationships in the intercorrelation matrices involving differences in the proportions shows more temporal stability than using the ratios of proportions.

This claim has received support from Swanson (1978) who evaluated the two regression-symptomatic procedures in the context of estimating small, highly concentrated sub-state populations. In addition, he reported that specific to such small domains the ratio-correlation method may suffer from a "destruction" of information; whenever P_{h1i} or P_{h2i} are zero, taking differences rather than ratios retains information that ratios destroy.

Morrison and Relles (1975) present another method that is similar to the ratio-correlation approach, but it uses a logarithmic form which, they claim, results in a dependent variable that is less volatile and more symmetric, and therefore more in accordance with the assumptions justifying least squares.

5. *Sample-Regression Method*

The sample-regression method is based on a regression equation using selected symptomatic indicator variables, measured for each domain, as independent variables, but current sample data for the variable of interest as the dependent variable. This method relies heavily on the current data at hand to determine the parameters in the model. In fact, the data used are exclusively postcensal and the equation that is produced is not constrained by prior logical assumptions as, for example, in the ratio-correlation approach. A study reported by Hansen *et al.* (1953) used double sampling in a regression relationship for improving local estimates, but they did not use symptomatic data nor a multivariate approach. Woodruff (1966) later provided an extension. However the approach as it is known today is generally attributed to Ericksen (1971), since he extended it to the multivariate case.

The method is similar to the ratio-correlation procedure. The variables r_{hbi} are calculated as in (4.2) and sample estimates of the postcensal growth for the variable of interest, calculated at the primary sample unit (PSU) level, are obtained as ratios, which are given by $\hat{R}_{hb} = \hat{q}_{h3}/q_{h2}$, where $\hat{q}_{h3}$ is the current estimate of the variable of interest, which is assumed to be available for a sample of PSU's. The sample estimates of postcensal growth for the PSU's are then regressed on the symptomatic indicators in order to directly estimate, by multiple regression, the relationship among the variables for the postcensal period. That is,

$$\hat{R}_{hb} = \hat{\beta}_0 + \hat{\beta}_1 r_{hb1} + \ldots + \hat{\beta}_p r_{hbp}.$$

The values of the symptomatic indicators for the local areas are then substituted into the estimated regression equation, in order to derive the current estimates as in the ratio-correlation method.

In comparison with regressions using old census data (Section 4), the sample-regression method avoids the problem of changes in the structural relations. For this gain, however, it sacrifices losses from sampling variations. These usually arise both from selections of the PSU's (local areas), and selections within them. Hence the relative precisions of the two methods depends on the balance of the two kinds of variation: obsolescence versus sampling. These will be influenced by the nature of variables, the dynamics of the situation and by the size and quality of the samples.

Ericksen (1974a) derives an expression of the mean square error of the sample-regression estimates, based on the level of error in the sample PSU's, as

$$MSE(\hat{R}_{hb}) = [(n - p - 1)\sigma^2{}_u/n] + [(p + 1)\sigma^2{}_v/n],$$

where $\sigma^2{}_u$ is the between PSU variance unexplained by the predictor variables, $\sigma^2{}_v$ is the within PSU variance, n is the number of PSU's, and p is the number of symptomatic indicators.

The sample-regression method has been demonstrated to perform well for the estimation of county and state populations using 1970 census data on population growth (see Ericksen 1973). However, although the method has mainly been applied to population estimation, the concept of using current samples together with symptomatic data is bound to be useful in many situations. It can also be adapted to non-linear multivariate situations.

6. *Synthetic-Regression Procedures*

The synthetic approach discussed in Section 3 is clearly limited by its ability to properly account for changes, since the time of their collection, in the distribution of the associated variables, across the small domains. For example, the introduction of a new factory or industry into a particular small region will have obvious implications on the work force in that area. The building of a retirement center would similarily lead to a change in the proportion of old aged in the region and therefore public health requirements. One way to help overcome this lack of sensitivity to local area changes that are not reflected in the large domains, is to introduce additional symptomatic variables by way of a regression relationship. The sorts of indicators which could be used here are births, deaths, school enrollments, registered unemployment and so on.

Recognizing the limitation of the synthetic estimator to properly account for local factors, Levy (1971) proposed the ***regression adjusted synthetic method*** which uses symptomatic information at the local level in conjunction with the synthetic estimate. He considered the following model:

$$x_h{}^* = \alpha + \beta Y_h + \epsilon_h, \tag{6.1}$$

where $x_h{}^* = \{(X_h - \hat{x}_h)/\hat{x}_h\}100$, $\hat{x}_h$ is the synthetic estimate, X_h is the true value, Y_h is the value of the symptomatic variable, α and β are the regression coefficients to be estimated, and ϵ_h is a random error term. Were estimates $\hat{\alpha}$ of α and $\hat{\beta}$ of β available and the error term omitted, an estimator $\hat{x}_h{}^*$ of X_h could be derived directly from (6.1) as $\hat{x}_h{}^* = \hat{x}_h[(\hat{\alpha} + \hat{\beta} Y_h)/100 + 1]$.

Since $x_h{}^*$ is a function of the true value X_h (which is unknown), a different method must be used to estimate the linear coefficients. Briefly, one such method is to estimate α and β by least squares after combining small domains to form strata, from which reasonable unbiased estimates of X_h can be obtained. This approach can obviously be extended to a multiple regression situation. Clearly, the main problem lies in effectively estimating the regression coefficients.

Recently, Nicholls (1977) carried out a detailed study into the possibility of applying a synthetic-regression approach to the estimation of local area populations in Australia. The proposed approach, best termed the *combined synthetic-regression method,* is basically the same as the sample-regression method except that the synthetic estimate is added as an additional independent variable. With appropriate choice of symptomatic variables, the technique showed an improvement over the synthetic estimates and the standard sample-regression estimates. Viewed from the point of view of the synthetic estimates, this improvement is brought about by a reduction in the bias while not seriously affecting the variance. More recently Gonzalez and Hoza (1978) reported encouraging results with this method for the calculation of small area estimates of unemployment. However, their study was limited to the 1970 census year and there is a need for further investigation of the methodology for intercensal years to determine the feasibility as well as the reliability of the suggested estimation procedures. As with the sample-regression method, this approach suffers from the fact that estimates must be available for a sample of small areas and the applicability of the method is generally limited to local area estimation.

The use of highly associated symptomatic variables, like births, deaths and school enrollments are likely to help with local population counts, but highly associated symptomatic variables may not be available for many of the variables in which we are interested. In these situations, and even in situations where good symptomatic information does exist, the sensitivity problem might best be remedied by identifying extreme cases and treating them separately. Gonzalez and Hoza (1978), for example, found improvements by excluding outliers from the regressions.

7. *Base Unit Method*

In looking for a method that could easily be used for variables other than population size, Kalsbeek (1973) proposed a method based on splitting up the small domains into smaller *base* units, then classifying each of these base units into one of k groups of base units for which estimates can be obtained from a survey sample. The small domain estimates are then formed by taking weighted combinations of these base estimates. The method, while not unlike synthetic estimation, is more restrictive in that it is practical only for small domains defined as local geographic areas.

The survey frame is divided into constituent geographic units, which are termed base units. These base units may be blocks, enumeration districts, counties, or other geographic units. Survey data are needed for a sample of these base units. On the basis of the symptomatic information and the information on the variable of interest for these sample base units, they are grouped into k homogeneous groups with a suitable clustering algorithm. The local areas of interest are also broken down into constituent base units, and each unit is then classified, by use of the symptomatic information available, into one of the k groups. The estimation procedure is then as follows. An estimate for each of the k groups of base units is formed by taking a weighted average of the estimates for the sample base units that constitute each group. That is,

$$\bar{x}'_g = \sum_{l=1}^{n_g} w_{gl}\bar{x}'_{gl}, \qquad g = 1, \ldots, k,$$

where n_g is the number of sample base units in the *gth* group of base units, w_{gl} is the weight and $\bar{x}'_{gl}$ is the sample estimate of the mean for the *lth* base unit in the *gth* group, and $\bar{x}'_g$ is the resulting weighted average (mean) for the *gth* group of base units.

The final estimate, $\bar{x}'_h$, is then given by taking a weighted average of the averages for the k groups, where the weights, θ_{gh}, represent the composition of the k groups in the local area h of interest. Thus,

$$\bar{x}'_h = \sum_{g=1}^{k} \theta_{gh}\bar{x}'_g.$$

The advantage of this method over the sample-regression method is that no special functional form is assumed involving the variable of interest and the symptomatic variables. However, the method does suffer from the fact that estimates must be available for a sample of base units, which does restrict the application of the method, resulting in it being generally only applicable to local-area estimation. The base unit estimator is also biased and Cohen and Kalsbeek (1977) go part of the way to derive an approximate expression of its mean square error, under some constraints. More recently, Cohen (1978) has studied a modification of the base unit method, where the base units are grouped via methods along the lines of minimum variance stratification as opposed to clustering algorithms.

8. *Brief Appraisal of Existing Techniques*

Potential users must note the different data requirements of diverse methods; data availabilities may dictate the choice of methodology. Nevertheless, evaluation studies may help a choice between methods, when choice is possible.

In relation to total population estimation for local areas, there exist a growing number of evaluation studies comparing the various symptomatic approaches. Both Ericksen (1971) and Kalsbeek (1973) provide fairly up-to-date reviews of these evaluation studies and summarize their main findings. Given reasonable model choice, situations of fairly stable growth and good auxiliary data, the regression-symptomatic and sample-regression methods have been demonstrated to result in somewhat more accurate estimates of small area population growth than the symptomatic accounting techniques (see Ericksen 1974b, O'Hare 1976). Ericksen's sample-regression method has consistently performed the best in these studies. The administrative data records method (Section 2), however, was not included in these evaluations, but it has shown considerable potential especially for very small local areas.

In a more general setting, the regression-symptomatic and sample-regression methods are especially suited to the estimation for continuous variables, such as income. Of these methods, the sample-regression method seems to have the greatest potential and accuracy, whenever good sample data on the variable of interest are available for a sample of the small domains. As a result, it is largely restricted to local area estimation and is less adapted to the estimation for crossclasses. If current sample data are not available, then the ratio or difference correlation approaches can still be applied as long as the variable of interest was collected in the previous censuses. The user is warned, however, that the correlation approaches assume that the relationship between the variables established in the intercensal period carries over to the postcensal period; this may be adequate for local population estimation but may be too restrictive for many other variables.

The base unit method has been extensively compared to the sample-regression method by Kalsbeek (1973). It was found that the proposed method performed fairly well in linear settings, although the sample-regression method performed slightly better and would be recommended in these circumstances. Cohen (1978) compared these two methods in nonlinear settings, but restricted attention to fitting linear regressions to the nonlinear data. It should be pointed out that Ericksen's sample-regression method does not inherently restrict us to using a linear regression formulation; in some situations one should fit nonlinear models. The sample data, which is assumed to be available for both these methods, can be used to suggest the appropriate functional relationship. Consider also that the base unit method is more difficult to implement, since one first has to split the small domains into the smaller base units needed for the method. Therefore, the sample-regression method would appear to offer a greater generality, although additional investigation is warranted.

The synthetic method seems the most popular, versatile and simplest approach for frequency type variables. While there are no explicit model assumptions, the synthetic method does implicitly assume the stability of various interactions between the variable of interest and the variables used to define the subgroups, at the level of the ratio adjustments. Where these implicit assumptions are not valid, bias becomes a problem. More work needs to be done in identifying the underlying structure-preserving properties of the synthetic estimates and the resulting implications.

There seems little point in making extensive comparisons of the synthetic and sample-regression methods since the two methods have somewhat different data requirements. Instead, research should be concentrated on looking for a combination of these two methods, to obtain the strength of both. Such research has produced the combined synthetic-regression method, which may have the greatest potential of all the methods currently available. However, it does require sample data on the variable of interest to be available for a sample of small domains and considerable resources to implement. Further investigations of this approach are needed along the lines indicated by Gonzalez and Hoza (1978).

Since no single method is likely to be best for all small domain estimation problems, we should move towards a combination of methods so that the strengths of each can be utilized. However, there is also a need for better understanding the diverse methods, especially the reasons for and circumstances of their successes and failures, if such composite approaches are to be anything but guesswork. These aspects are reflected in the direction of the current research that is discussed in the following section.

As a final point, the level and quality of available data, for the variable of interest and for associated variables, essentially dictate the choice and accuracy of the existing techniques for small domain estimation. What we get out of the techniques is a direct function of what we put into them.

9. *Current Research and Lines for Future Advances*

As we have already seen, there are many approaches to the problem of small domain estimation and this diversity has carried over into current research. Some of these involve variations of methods already described, yet it is informative to indicate the directions they are taking.

Many of these new approaches are best identified as composite approaches, combining two or more separate techniques, as discussed in Sub-sections 9.1 through 9.4. The idea of a composite estimator for small domain estimation′is not new; it has been tried for symptomatic accounting techniques, and also advocated by the National Center for Health Statistics (1968). The synthetic-regression approaches are further examples of composite approaches that have seen some application.

9.1. *Composite Synthetic Approach*

Where sample estimates from the small domains of interest are available it may be advantageous to combine them with synthetic estimates. From the low variance of the biased synthetic estimate and the high variance of the direct estimate, a suitable combination will give values for both the variance and the bias between those of the two estimates. The problem is to select appropriate weights for the combination.

One approach, by Schaible *et al.* (1977), arrives at a composite estimate by weighting each component in proportion to the inverse of its squared error. Assume that the expected mean square error of the direct estimator is of the form b/n_h, and that of the synthetic estimator is b', where b and b' are constants and n_h is the sample size in the *hth* small domain. Then the *composite synthetic* estimator is given by $\tilde{x}_h{}^* = c_h x'_h + (1 - c_h)\hat{x}_h$, where $c_h = n_h/(n_h + b/b')$. The quantity b/b' is the small domain sample size at which the expected errors of the synthetic and direct estimators are equal. This approach is similar to the James-Stein approach discussed later in Sub-section 9.3, with a difference in the method of choosing weights. Schaible *et al.* (1977) showed that in estimating both the unemployment rates for county groups in Texas and the percent of the population completing college for states, the composite estimator had a MSE approximately 30% less than that of the synthetic estimator. The preliminary results also indicate that the composite estimator is remarkably robust against poor estimates of the unknown quantity b/b'. Investigations of the properties of this composite estimator are continuing, and Schaible (1978) reports on an evaluation of the choice of weights for the composite synthetic approach.

9.2. *Composite Ratio-Correlation and Sample-Regression Method*

The ratio-correlation and sample-regression methods are not so much different methods but different estimates; differences between the two arise chiefly from different assumptions regarding available data. The sample-regression method uses current sample information to estimate the regression coefficients, while the ratio-correlation technique uses more precise but out of date coefficients.

Royall (1974) suggests the possibility of using a particular linear combination of estimates of the old and new coefficients. We can extend this further by including the synthetic estimator as another explanatory variable, but these ideas still need exploration.

9.3. *James-Stein and Bayesian Estimates*

Considerable interest has recently been shown in the possibility of applying a James-Stein procedure to small domain estimation problems. Space permits only a brief discussion of the

James-Stein estimator and its descendants, but interested readers can refer to Efron and Morris (1973). This empirical Bayes approach parallels the classical Bayes approach by Box and Tiao (1973), Chapter 7. Suppose we have sample estimates, x'_h, for H different small domains, each of which have equal variances, ν, and different means, u_h. Given a set of prior estimates, p_h, a logical estimator of u_h is

$$\tilde{x}_h = cx'_h + (1 - c)p_h, \tag{9.1}$$

where $(0 \leq c \leq 1)$. The prior estimates, p_h, could include the overall mean, or the means of large domains; but they can also be synthetic estimates, sample-regression estimates, and so on. Now, for fixed c, the expected squared error of $\tilde{x}_h$ is given by

$$R(u_h, \tilde{x}_h) = \sum_h E_{u_h}(u_h - \tilde{x}_h)^2,$$

which is minimized by choosing $c = w/(w + v)$, where $w = \sum_h(p_h - u_h)^2/H$.

Substituting this value of c in (9.1) results in Min $R(u_h, \tilde{x}_h) = cHv$, but it is also seen that $R(u_h, x'_h) = Hv$. Thus Min $R(u_h, \tilde{x}_h) \leq R(u_h, x'_h)$. Based on this result, the James-Stein estimator is simply (9.1) with c estimated from the sample as $\hat{c} = 1 - (H - 2)v/s$, where $s = \sum_h(x'_h - p_h)^2$. For $H \geq 3$, the James-Stein estimator has risk (expected squared error) less than that of x'_h for all u_h.

Note that this method can be extended to situations where the x'_h have unequal variances, v_h, of which the composite-synthetic estimator (Sub-section 9.1) is a special case. It can also be used in situations where only samples are available, without auxiliary data from either censuses or registers. It, and the double sampling-regression method described by Hansen *et al.* (1953), are thus available where the others are not. But it can also be combined with auxiliary information. Recently, the U.S. Bureau of the Census has applied a modified James-Stein estimator to sample data from the 1970 Census in order to estimate base figures for small areas in the Census Bureau's program of estimation for the purposes of General Revenue Sharing (Fay and Harriot 1977, Fay 1978). The prior estimate, p_h, considered in this study was the sample-regression estimate. The work in this area is still in its early stages, but it can reasonably be expected that James-Stein procedures will play an important role in small domain estimation.

9.4. *Prediction Approach*

In the prediction approach, a super-population probability model of the relationship between the variable of interest and symptomatic variables is assumed and from this are derived "optimal" sub-domain predictors. Estimators of the small area characteristics are constructed by assuming, for example, a linear regression relationship between the variable of interest and symptomatic auxiliary variables. Following the work of Royall (1970), the best predictor can then be expressed as a linear combination of the observed values of the sampled units and of a predictor of the unobserved population units.

This approach to small domain estimation has been investigated by Laake (1977), Royall (1977), Royall (1978), and Holt, Smith and Tomberlin (1977). One advantage is that it generally yields estimates of MSEs, under the model, as measures of reliability. Comparisons of the resulting predictor with the "conventional" synthetic estimator have been carried out by Laake (1977).

9.5. *A Categorical Data Analysis Approach*

Structuring of the small domain estimation problem within a categorical data analysis framework appears as a logical approach, but it has received little attention to date.

Recently, however, Freeman and Koch (1976) made reference to the applicability of marginal adjustment (raking) of contingency tables for local area estimation. Essentially, they considered the restrictive case where "inaccurate" small domain estimates are available, derived usually from a survey. The method then adjusts these estimates to agree with a known (accurate) set of estimates at higher levels of aggregation.

At the same time the Australian Bureau of Statistics was investigating a related approach to a different and more usual situation brought about by data availabilities, as reported by Chambers and Feeney (1977). Small domain estimates are assumed available from some previous source (usually the census) and survey data provides current estimates at higher levels of aggregation. The basic feature of this approach is the assumption of the stability (in some sense) of the *association structure:* the structure inherent in the frequencies recorded at the small domain level for the variable of interest, cross-tabulated by some associated variables at some previous time, usually in the census. As with the synthetic method, current sample information is assumed to be available at higher levels of aggregation. This current information specifies new margins in the cross-tabulated data and is usually referred to as the *allocation structure.* The estimation process uses an iterative proportional fitting (IPF) algorithm, as first described by Deming and Stephan (1940), to force the original cross-tabulation, as established at the previous census, to agree with the new margins. Small domain estimates can then be obtained by summing the resulting adjusted cross-tabulation over the appropriate cells. A complete and structured treatment of this approach, and extensions of it aimed at improving its efficiency, are given by Purcell (1979), together with a computer program for its implementation.

One feature of this approach is the implicit assumption of an underlying super-population model governing the behavior of the small domain frequencies over time. In this context, it is usual to assume that the data follow either a Poisson or multinomial model. The resulting IPF estimates have the property that they preserve all the interactions specified by the *association structure,* except those respecified by the *allocation structure.* In addition, the estimates maximize the likelihood equation of the multinomial distribution. A further consequence is that the synthetic estimates, as specified in equation (3.1), are equal to the first approximation of the IPF solution, and are optimal (best asymptotically normal) under appropriate constraints.

The main advantages of this approach are that it is extremely flexible, and the properties of the resulting estimates can be made more apparent. The flexibility is reflected in the fact that it is applicable to all small domains and does not require estimates for a sample of small domains to be available, as several of the alternative approaches do. In addition it is particularly conducive to the use of nominal and qualitative variables, which occur frequently in this area. It also helps to put a logical framework on the small domain problem.

The main disadvantage of the procedure is that it is only applicable to situations where the variable of interest can be represented by frequencies. Thus this approach does not address the problem of estimating small domain non-frequency characteristics such as average income, average expenditure and so on. Such nonfrequency variables have a continuous distribution and are more suited to "conventional" estimation approaches such as regression.

Chambers and Feeney (1977) discuss an application of this approach to the estimation of small area estimates of work force status. Bousfield (1977) has applied a similar approach to the estimation of intercensal estimates of age by sex by race for the Chicago SMSA. Finally, in an extensive empirical evaluation of "synthetic" state estimates of mortality due to each of four different causes, Purcell (1979) has demonstrated that the estimators resulting out of this approach significantly outperform the basic synthetic estimator, given in (3.1). The evaluation involved a study of the performance of the alternate estimators over the full 10-year postcensal period, on yearly data, 1960 through 1970.

10. Summary Comments

While there exist several alternative approaches to the small domain problem, the methodology still lacks a consistent logical approach. But aiming for this may be unproductive since the best approach may be problem specific, not amenable to a single formal characterization. Instead we may aim at a categorization of estimation situations and of variables of interest, to serve as a framework for choosing the most reasonable small domain estimation method for the situation at hand.

The central issue is the appropriateness of the underlying models, obtained either implicitly through the basic assumptions on which the techniques are based, or directly through our assumptions about the data structure. The choice therefore, is not one of methods, but really one of models, with the success of small domain estimation depending largely on being able to identify symptomatic variables that are highly related to the variable of interest, and on understanding the *association structure* existing between the variables.

We view recent developments of estimation for small domains as a significant enlargement of the scope of statistics beyond its past concentration either on overall statistics or on individual predictions. With the increasing data requirements of our society, it is clear that small domain estimation will continue to grow in importance. Further developments will expand theory and applications to new problems and to new variables. Better ways will be found to combine the strengths of the diverse sources of data—samples, censuses and registers—to construct synthetic small domain data bases. Estimates derived from these data bases will be simultaneously more detailed, timely, and accurate than those currently available.

Résumé

On peut obtenir des données de santé, économiques et sociales, complètes et à des instants donnés, à partir d'échantillons en général seulement pour des aires géographiques importantes et de grands sous-groupes de la population. On peut à partir de recensements avoir des estimations sur des petits domaines, mais rarement et seulement alors pour peu de variables. Les plans des services de santé et autres activités gouvernementales ne peuvent dépendre des sources traditionnelles de données. Les données doivent être plus courantes et plus complètes que celles que fournissent ces sources.

Comme on a besoin d'estimations pour une grande variété de domaines, l' article présente une définition et une classification des domaines pour clarifier la direction de cette revue. Les techniques d'estimation existant pour les petits domaines sont divisées en différentes approches et revues séparément. On présente les méthodologies de base de ces techniques en même temps que les données qu'elles nécessitent et leurs limitations. On évalue brièvement les techniques existantes en fonction de leurs performances et de leur potentiel d'application. On passe aussi en revue les approches courantes de recherche et on indique des directions possibles de futurs développements.

References

Bogue, D. J. (1950). A technique for making extensive population estimates. *Journal of the American Statistical Association 45,* 149–163.

Bogue, D. J. and Duncan, B. D. (1959). A composite method for estimating postcensal population of small areas by age, sex and colour. National Office of Vital Statistics, *Vital Statistics—Special Reports 47,* No. 6.

Bousfield, M. V. (1977). Intercensal estimation using a current sample and census data. *Review of Public Data Use 5, No. 6,* 6–15.

Box, E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison Wesley, Cambridge, Massachusetts.

Chambers, R. L. and Feeney, G. A. (1977). Log linear models for small area estimation. Paper presented at the Joint Conference of the CSIRO Division of Mathematics and Statistics and the Australian Region of the Biometrics Society, Newcastle, Australia, 29 August–2 September. *Biometrics* Abstract #2655.

Cohen, S. B. (1978). A modified approach to small area estimation. Unpublished Ph.D. thesis, University of North Carolina, Chapel Hill, North Carolina.

Cohen, S. B. and Kalsbeek, W. D. (1977). An alternative strategy for estimating the parameters of local areas. *1977 Proceedings of the Social Statistics Section,* American Statistical Association, 781–785.

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 11,* 427–444.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association 68,* 117–130.

Ericksen, E. P. (1971). A method for combining sample survey data and symptomatic indicators to obtain estimates for local areas. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Ericksen, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography 10,* 137–160.

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

Ericksen, E. P. (1974b). Developments in statistical estimation for local areas. *Census Tract Papers, Series GE 40, No. 10.* U.S. Bureau of the Census, U.S. Government Printing Office, Washington, D.C.

Fay, R. E. (1978). Some recent Census Bureau applications of regression techniques to estimation. Presented at the NIDA/NCHS Workshop on Synthetic Estimates, Princeton, New Jersey, April 13–14. Proceedings to be published by NIDA, 1979.

Fay, R. E. and Herriot, R. (1977). Estimates of income for small places: An application of James-Stein procedures to census data. Submitted to the *Journal of the American Statistical Association.*

Freeman, D. H. and Koch, G. G. (1976). An asymptotic covariance structure for testing hypotheses on raked contingency tables from complex sample surveys. *1976 Proceedings of the Social Statistics Section,* American Statistical Association, 330–335.

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

Ghangurde, P. D. and Singh, M. P. (1978). Evaluation of efficiency of synthetic estimates. To appear in *1978 Proceedings of the Social Statistics Section,* American Statistical Association.

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.

Gonzalez, M. E. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria. 18–25 August.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory I.* John Wiley and Sons, Inc., New York.

Holt, D., Smith, T. M. F. and Tomberlin, T. J. (1977). Synthetic estimation for small sub-groups of a population. Unpublished technical report, University of Southhampton, England.

Kalsbeek, W. D. (1973). A method for obtaining local postcensal estimates for several types of variables. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Kish, L. (1965). *Survey Sampling.* John Wiley and Sons, Inc., New York.

Laake, P. (1977). A prediction approach to sub-domain estimation in infinite populations. Central Bureau of Statistics of Norway (Mimeographed).

Laake, P. (1978). An evaluation of synthetic estimates of employment. *Scandinavian Journal of Statistics 5,* 57–60.

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, Artical 88.*

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

Martin, J. H. and Serow, W. J. (1978). Estimating demographic characteristics using the ratio-correlation method. *Demography 15,* 223–233.

Morrison, P. A. and Relles, D. A. (1975). A method for monitoring small area population changes in cities. *Review of Public Data Use 3, No. 2,* 10–15.

Namboodiri, N. K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography 9,* 443–453.

Namekata, T. (1974). Synthetic state estimates of work disability. Unpublished Ph.D. thesis, University of Illinois, Champagne, Illinois.

Namekata, T., Levy, P. S. and O'Rourke, T. W. (1975). Synthetic estimates of work loss disability for each state and the District of Columbia. *Public Health Reports 90,* 532–538.

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.

National Center for Health Statistics (1977a). *State Estimates of Disability and Utilization of Medical Services, United States, 1969–1971.* D.H.E.W. Publication No. (HRA) 77-1241. U.S. Government Printing Office, Washington, D. C.

National Center for Health Statistics (1977b). *Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey.* D.H.E.W. Publication No. (P.H.S.) 78-1349. U.S. Government Printing Office, Washington, D. C.

Nicholls, A. (1977). A regression approach to small area estimation. Australian Bureau of Statistics, Canberra, Australia, March (Mimeographed).

O'Hare, W. (1976). Report on a multiple regression method for making population estimates. *Demography 13,* 369–379.

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

Pursell, D. E. (1970). Improving population estimates with the use of dummy variables. *Demography 7,* 87–91.

Rives, N. W. (1976). A modified housing unit method for small area population estimation. *1976 Proceedings of the Social Statistics Section,* American Statistical Association, 717–720.

Rosenberg, H. (1968). Improving current population estimates through stratification. *Land Economics 44,* 331–338.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika 57,* 377–387.

Royall, R. M. (1974). Discussion of papers by Gonzalez and Ericksen. *Census Tract Papers, Series GE 40, No. 10.* U.S. Bureau of the Census, U.S. Government Printing Office, Washington, D. C.

Royall, R. M. (1977). Statistical theory of small area estimates—use of predictor models. Unpublished technical report prepared under contract from the National Center for Health Statistics.

Royall, R. M. (1978). Prediction models in small area estimation. Presented at NIDA/NCHS Workshop on Synthetic Estimates, Princeton, New Jersey, April 13–14. Proceedings to be published by NIDA, 1979.

Schaible, W. L. (1978). Choosing weights for composite estimates for small area statistics. To appear in *1978 Proceedings of the Social Statistics Section,* American Statistical Association.

Schaible, W. L., Brock, D. B. and Schnack, G. A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *1977 Proceedings of the Social Statistics Section,* American Statistical Association, 1017–1021.

Schmitt, R. C. and Crosetti, A. H. (1954). Accuracy of the ratio-correlation method of estimating postcensal population. *Land Economics 30,* 279–280.

Snow, E. C. (1911). The application of the method of multiple correlation to the estimation of post-censal populations. *Journal of the Royal Statistical Society 74,* 575–620.

Starsinic, D. E. (1974). Development of population estimates for revenue sharing areas. *Census Tract Papers, Series GE 40, No. 10.* U.S. Bureau of the Census, U.S. Government Printing Office, Washington, D. C.

Swanson, D. A. (1978). An evaluation of "ratio" and "difference" regression methods for estimating small, highly concentrated populations: The case of ethnic groups. *Review of Public Data Use 6, No. 4,* 18–27.

United Nations (1950). *The Preparation of Sampling Survey Reports.* Statistical Papers, Series C, No. 1.

U.S. Bureau of the Census (1947). *Population-Special Reports, Series P-47, No. 4,* U.S. Government Printing Office, Washington, D. C.

U.S. Bureau of the Census (1949). Illustrative examples of two methods of estimating the current population of small areas. *Current Population Reports, Series P-25, No. 20.* U.S. Government Printing Office, Washington, D. C.

U.S. Bureau of the Census (1966). Methods of population estimation: Part I, Illustrative procedure of the Census Bureau's component method II. *Current Population Reports, Series P-25, No. 339.* U.S. Government Printing Office, Washington, D. C.

U.S. Bureau of the Census (1969). Estimates of the population of counties and metropolitan areas, July 1, 1966: A summary report. *Current Population Reports, Series P-25, No. 427.* U.S. Government Printing Office, Washington, D. C.

U.S. Bureau of the Census (1975a). Population estimates and projection. *Current Population Reports, Series P-25, No. 580.* U.S. Government Printing Office, Washington, D. C.

U.S. Bureau of the Census (1975b). *The Methods and Materials of Demography.* By Henry S. Shryock, Jacob S. Siegel, and Associates. Third Printing, U.S. Government Printing Office, Washington, D. C.

Woodruff, R. S. (1966). Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade. *Journal of the American Statistical Association 61,* 496–504.

Received September 1978; *Revised December* 1978