International Statistical Review, 48 (1980) 3–18 Longman Group Limited/Printed in Great Britain

Postcensal Estimates for Local Areas (or Domains)

Noel J. Purcell[1] and Leslie Kish[2]

[1] *Australian Bureau of Statistics; and* [2] *University of Michigan, Ann Arbor*

Summary

Demands are growing for more timely and varied statistics for administrative units and for other small domains. These demands exist in several countries now, and they are bound to spread to all others as useful methods become known. Possibilities depend on the availability of data and of other resources, but they are also being broadened with the introduction of newer methods. We present a review and classification of those methods together with an extensive bibliography. Then we describe a new method; iterative proportional fitting (IPF), which fits sample data with iterative techniques to a flexible nonlinear model that preserves specified relationships found in an earlier census. The model and method are justified within a broad theoretical framework. Great improvements over current methods are shown in empirical results, and are promised for many situations.

1 Statement of the problem

Timely and complete social and economic data can be obtained from national sample surveys, but usually only for *major domains* of study. Domains can be local areas, often administrative units, such as geographic areas, for which separate estimates are planned, and which also tend to be partitions in the collection process, whether of censuses or of sample surveys. Or, on the contrary, they can be 'crossclasses' of the population and of the sample, which cut across the partitions of the collection and the sample design; for example, age and sex categories. Less commonly used than either of the above are domains that have not been distinguished in the sample selection, but tend to concentrate unevenly in the primary sample units. Estimates for *small domains* – any 'small' subclass or subdivision of the original domains of study – are generally unavailable from typical samples, and are obtained from population censuses or from administrative registers, and sometimes from special-purpose surveys. However, effective planning of social services and other activities cannot depend on these traditional data sources; the data must be more current, more complete, and more relevant than these sources provide. Estimates for local areas as administrative units appear as the most salient and common concern for detailed data, but crossclasses and other small domains can also be important. Several of these methods are equally pertinent to both.

Since the sizes of the small domains influence the choice and applicability of methods, a classification of domains, based on their sizes, is presented here (Purcell and Kish, 1979). It can remind us of the practical differences between the types of domains and help us avoid the common mistake of considering 'statistics for domains' as one homogeneous problem. The boundaries of this classification are stated very roughly to orders of magnitude and should not be taken too seriously; they depend on the variables and the statistics estimated, on the sizes of samples and of populations, on the precisions and decisions involved, etc.

1. *Major domains*, composing 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 the unemployed or the disabled.

3. *Mini domains*, comprising from 1/100 to 1/10,000 of the population. Examples: populations of counties (more than 3000 of them in the U.S.A.), a two-fold classification like state by work force status, or a three-fold classification like age by occupation by education.
4. *Rare domains*, comprising less than 1/10,000 of the population. Examples: populations of small local areas, perhaps all classified by various ethnic groups.

For major domains, standard estimates, basically without bias, are generally available from probability methods of survey sampling. However, frequently for minor domains and usually for mini domains, the standard methods of survey estimation break down, because the sample bases are ordinarily too small for any useable reliability, and new methods are needed. For rare domains, sample surveys are usually useless; separate and distinct methods are required (see for example, Kish, Section 11.4). Thus this discussion is directed principally to classes 2 and 3, and we use the general term small domains to refer to these two classes throughout this paper.

2 Existing local area estimation techniques

2.1 *Symptomatic Accounting Techniques* (*SAT*)

These techniques utilize current data from administrative registers in combination with their statistical relationships established with earlier census data. A series of techniques using diverse registers have been developed and tested by the U.S. Bureau of the Census, and several good reviews of them are available (U.S. Bureau of the Census, 1975*a* and *b*; Ericksen, 1971; Kalsbeek, 1973). Most of these techniques deal with small area estimates of the total population, though other estimates could similarly be designed with appropriate data and techniques.

The diverse registration data used in the U.S.A. have included the numbers and rates of births and deaths, of existing and new housing units, of school enrolments, and of income tax returns. These 'symptomatic' variables need not concern births and deaths directly, nor would these account for the large effects of migration. But fluctuations in the symptomatic variables should be highly related to changes in population totals or in its components. Thus school enrolments are highly related to numbers of children and of young parents, and death rates to older persons. Diverse techniques have been described; Census Component Methods I and II (U.S. Bureau of the Census, 1949, 1966); the vital rates technique and the composite method (Bogue, 1950; Bogue and Duncan, 1959); the housing unit method (U.S. Bureau of the Census, 1969); and the administrative data records method (Starsinic, 1974).

Clearly the relative advantages of diverse techniques will depend greatly on the nature of the desired estimates, and especially on the kinds of registration data which are available in each country.

2.2 *Synthetic estimation*

This name has been used for a form of ratio estimation which combines recent sample means, $\bar{x}'_{\cdot g}$, for subclasses (g) at the large domain level with Census proportions $Y_{hg}/Y_{h\cdot}$ of the subclasses in the small areas h to obtain estimated means $\bar{x}_{h\cdot}$ for small areas (domains):

$$\bar{x}_{h\cdot} = \sum_{g} (Y_{hg}/Y_{h\cdot})\bar{x}'_{\cdot g}.$$

The sample size is presumed to yield useful estimates of the $\bar{x}'_{\cdot g}$ because these are reasonably large and widespread, but not of the $\bar{x}'_{hg}$ nor of the $\bar{x}'_{h\cdot}$. The estimates $\bar{x}_{h\cdot}$ can refer to the general population or to some subset, related (preferably highly and linearly) to the auxiliary statistics $Y_{hg}/Y_{h.}$. These statistics can refer to the general population in the Census, or to some other source and to other populations.

For estimates of total persons and of unemployment see Gonzalez (1973); Gonzalez and Waksberg (1973); Schaible *et al.* (1977); Gonzalez and Hoza (1978); also work from Norway by Laake and Langva (1976) and by Laake (1978).

For estimates in the health field see National Center for Health Statistics (NCHS) (1968, 1977*a* and *b*); Levy (1971); Nameketa (1974); Nameketa *et al.* (1975).

A different form of the synthetic estimator is $\hat{x}_{h.} = \sum_g \hat{x}_{hg} = \sum_g (Y_{hg}/Y_{.g})x'_{.g}$, where $x'_{.g}$, are sample estimates for subclasses g. This has been proposed and investigated in Australia by Purcell and Linacre (1976), and in Canada by Ghangurde and Singh (1977).

2.3 *Regression – symptomatic techniques*

These methods use multiple linear regressions computed on ratios of change between two Census base periods ($t = 1, 2$) and between the predicted ratio of change, R_{hi}, and the predictors r_{hi} ($i = 1, 2, ..., p$). The p predictor ratios are based on 'symptomatic' variables, which are available from administrative registers for local areas, not only for Census years but also currently for postcensal periods.

For each symptomatic variable (i) the *proportion* for area h, $P_{h1i} = Y_{h1i}/Y_{.1i} = Y_{h1i}/\sum_h Y_{h1i}$ is computed for Census period 1. A similar proportion is computed as $P_{h2i} = Y_{h2i}/Y_{.2i}$ for Census period 2. The ratio of change $r_{hi} = P_{h2i}/P_{h1i}$ for the proportions in area h also equals $(Y_{h2i}/Y_{h1i})/(Y_{.2i}/Y_{.1i})$ the proportion of the change ratio which belongs to the hth component area. A similar ratio R_h is computed for the predicted variable. Then the multiple regression

$$R_h = B_0 + B_1 r_{h1} + B_2 r_{h2} + ... + B_p r_{hp}$$

is computed to obtain the regression coefficient B_0, B_1, B_2, ..., B_p, linking changes between the two Census periods in the predicted population ratio of change R_h to those in the p predictor changes r_{hi}.

Then the ratio of change R'_h in the local area h occurring in the postcensal period (between periods 2 and 3) is predicted from the computed ratios of change $r'_{hi} = P_{h3i}/P_{h2i}$ in the symptomatic variables. For this equation we must use the coefficients B_i obtained from the relations observed between the earlier intercensal periods 1 and 2:

$$R'_h = B_0 + B_1 r'_{h1} + B_2 r'_{h2} + ... + B_p r'_{hp}.$$

The ratio-correlation techniques follow a paper by Schmitt and Crosetti (1954). References and modification (averaging of univariate estimates) are given by Namboodiri (1972). Stratification was investigated by Rosenberg (1968) and dummy variables by Pursell (1970). Martin and Serow (1978) applied the method to estimate age and sex compositions in local areas. O'Hare (1976) investigated the advantages of differences, $d_{hi} = P_{h2i} - P_{h1i}$, in the place of ratios, r_{hi}, and was supported by Swanson (1978), whereas Morrison and Relles (1975) used a logarithmic form.

2.4 *Sample – regression methods*

The regression method above uses the coefficients B_i from structural relations established from earlier Census periods, but changes in those relations bring errors into the current postcensal estimates. These biases of obsolescence in the B_i can be overcome by computing b_i from large current samples, but at the cost of sampling errors in those b_i. The ratios r'_{hi} are computed from postcensal changes in symptomatic variables for all local areas, and the R_h are computed only for each local area in the sample. Then from

$$R_h = b_0 + b_1 r'_{h1} + b_2 r'_{h2} + ... + b_p r'_{hp}$$

the regression coefficients b_i are estimated from the local areas in the sample. These values of b_i are then used with all values r'_{hi} to estimate R'_h for all local areas in the population, not only in the sample.

This method was investigated by Ericksen (1971, 1973, 1974); in his studies the gains from avoiding obsolescence slightly overcame losses from sampling. This balance depends on the size and quality of the samples, on the dynamics of the structural model, and on the value of the variables. The computed values of the b_i are subject to sampling variances which have two components: a between area component when only a sample of areas is sampled, and a within area component due to sampling within areas (Ericksen, 1974).

Combinations of the method with other regressions and with synthetic estimators are possible. Nicholls (1977) added the synthetic estimate as an additional independent variable, and Gonzalez and Hoza (1978) also reported on improvements over synthetic estimates with this combination.

In the above, the same units (counties in the USA) were supposedly used both for primary selections (PSU's) and for local area estimation. But Kalsbeek (1973) proposed that greater precision may be had with a *base unit method*, which utilizes the secondary (or later) selection units. Furthermore, he avoided linear regression for greater flexibility with a suitable clustering algorithm to sort the base units into a set of homogeneous groups. Later Cohen (1978) used minimum variance stratification for homogeneous grouping.

2.5 *Stein–James, Bayesian and 'shrinkage' estimates*

Here we note jointly several kinds of estimates which all involve weighted means of the sample estimates x'_h for H domains combined with an auxiliary estimator p_h in the form

$$\bar{x}_h = Cx'_h + (1-C)p_h,$$

where $0 \leqq C \leqq 1$.

The auxiliary p_h may be the overall mean $\bar{x}$ of the sample, and then the method is based on data from the sample alone, and needs no auxiliary data. This can be an advantage for some situations (some countries and some kinds of statistics) where auxiliary data are judged to be lacking. Valid and useful choice of C presents difficult problems of balance between increased variances and basically unknown biases. Three lines of approach are 'shrinkage' methods (Thompson, 1968), 'empirical Bayes' methods (Efron and Morris, 1973), and 'Bayesian' methods (Box and Tiao, 1973, Ch. 7).

The auxiliary estimators p_h may also come from outside the sample. They may be synthetic or sample regression estimates, as in a study by the US Census Bureau (Fay and Herriot, 1977; Fay, 1978). In an early study a method of double sampling with regression was tried (Hansen *et al.*, 1953, Section II.5; Woodruff, 1966).

Other possibilities may be advanced for future investigations with this method of potentially great flexibility. Instead of the overall mean we can use the means of large domains for estimates of the small domains within them; a kind of stratified estimator. Furthermore, multifactor classification may be used in

$$\bar{x}_h = Cx'_h + \sum_i C_i p_{hi},$$

where each C_i identifies and weights a different 'predictor' of the variable $\bar{x}_h$, and

$$C + \sum_i C_i = 1.$$

3 Choice of methods in diverse situations

From the variety of available methods several lessons may be learned. First, that one may find among them a better method than the one he is arbitrarily using at present for small domains; this is often the passive 'null' method of continued reliance on the last decennial census which may be 12 years out-of-date.

Second, there is no single method that is best for all situations. Great differences between countries exist in the sources and quality of data available: the scope and quality of its census; the extents, contents and sizes of its sample surveys; and especially the scope and quality of its administrative registers. However, passive and negative attitudes are generally unjustified since every country has some resources, and ingenuity and effort can find unused resources of data. These may be of apparently different origins, but potentially useful because of high correlation with population sizes. It may be possible to find diverse data for different portions of the population; in 2.1 we note an example from the USA with school enrolments more useful for young age groups, and death registration for old ages. Perhaps in some countries one source is highly correlated with the rural population and another with the urban, or different states or provinces may have diverse sources.

Furthermore the choice of sources and methods should vary with and depend on the nature of the statistics, on the desired estimates, and also on the domains to which they pertain. Note also the effects of the lapse of time since the last census; this may be seen in Fig. 2 in divergences between estimates. More generally, the balance between the biases of a census and the variance of a large sample survey will move in favor of the latter during the 10 years between decennial censuses. The balance will also move in favor of less accurate but more current registers. The balance of sample surveys versus censuses or registers should depend on the sample size, but a fixed size sample has relative advantages in smaller populations. A general treatment of the relative advantages of censuses, of sample surveys and of administrative registers would be too complex and too long here but obviously the choice should depend on many factors and criteria (Kish, 1979).

Finally, the choice between methods is more difficult because the 'best' is often not clear even after the event. Errors of the estimates arise from biases chiefly, and 'true values' are usually not available for measuring the biases directly. Tests must be combinations of empirical and model bases, often depending eventually (and uncertainly) on results from decennial censuses. Better methods and criteria must be pursued with several methods and over the long range with an evolutionary approach and with patience.

The models of Section 4 should clarify the nature of the choice, and the new methods shown in Sections 5 and 6 should prove the 'best' in many situations, we believe.

4 A new approach: structure preserving estimation (SPREE)

Most small domain estimates concern variables represented by frequencies (count data), and these can readily be structured within a framework of categorical data analysis. Thus our framework is designed for and has been tested on frequency data. But it can also be adapted to estimate non-frequency characteristics (such as *per capita* income), as we demonstrate later in this section.

Categorical data analysis has been used for other purposes, but it is a new method for small domain estimation. A form of it appears in a study conducted at the Australian Bureau of Statistics, as initially reported on by Chambers and Feeney (1977); and Bousfield (1977) in a related context has proposed a similar approach. In a more unified treatment, Purcell (1979) fully documents this methodology and clearly develops and formulates the properties of the resulting estimates and their link to the ordinary, synthetic estimates.

This approach can utilize wide varieties and forms of information on variables that are associated with the estimand, i.e. the variable of interest. For example, age, sex and race are highly related to rates of employment of local areas. These *associated variables*, which effectively divide the population into crossclasses (cutting across local areas), will be seen to play an important role in the proposed estimation procedure.

Two explicit assumptions concerning data availabilities are made. First, we need an *association structure:* to establish at some previous time point (usually a census) the relationships between the estimand and associated variables, at the required small domain level. Second, we need an *allocation structure*: to establish current relationships between the variable of interest and the associated variables at the large domain level, accumulated over the small domains of interest. In addition, we may also utilize other current information on the variables, such as the small domain population sizes. The current information is typically obtained from large scale sample surveys, but other sources can be utilized. In general, the data constituting both the association and allocation structures are obtained directly from hard data sources (e.g., the census), but they may also be model based.

The information constituting the association structure is completely specified by the three-dimensional contingency table N with cell counts N_{hig} ($h = 1, \ldots, H, i = 1, \ldots, I, g = 1, \ldots, G$), where the subscripts refer respectively to the hth small domain, ith category of the estimand variable, and gth category of the associated variable dimension. The allocation structure (current information) actually represents updated margins for the association structure, N, and we denote the set of cell counts constituting these margins by $\boldsymbol{m}$. For example, if the allocation structure is defined by the margin of the estimand variable i and the associated variables g, then $\boldsymbol{m}$ is represented by the cell counts $m_{\cdot ig}$, where the dot subscript is used to denote summation over the subscript h for local areas.

Given the two data structures, N and $\boldsymbol{m}$, we first need updated estimates of the association structure, N, and we denote these estimates by $\boldsymbol{x}$, with cell counts x_{hig}. Our ultimate interest, however, is in the margin of $\boldsymbol{x}$ for small domains by the estimand variable; that is, in the estimated cell counts $x_{hi\cdot}$, which are obtained by simply summing over the associated variable dimension, g.

Example: For a graphical illustration only, consider the simple situation of estimating the number of employed and unemployed persons for each of two counties (local areas). Race is assumed to be highly associated with employment status and is used as a single associated dichotomous variable. Diagramatically then, our available information is represented as in Fig. 1. The association structure, N, in $2^3 = 8$ cells, represents the data obtained from the previous census, while the allocation structure, $\boldsymbol{m}$, given by the margin of employment status by race (summed over the counties in four cells) is derived from the current survey. In other words, we obtain current information on the back margin $m_{\cdot ig}$ in Fig. 1, but such data are not available, or are of unacceptable reliability at the county level.

The procedure then is to adjust the association structure (original table) to conform to the current data in the allocation structure (updated margin), while preserving (as much as possible in some way) the relationships between the variables in the association structure, N_{hig} (as established at the previous census). The resulting adjusted table x_{hig} is not stable and it is summed across the associated variable, g, to obtain the required estimates $x_{hi\cdot}$. In terms of Fig. 1, our required estimates are represented by the four cells on the right hand margin (the margin for employment status by county).

Extensions of the above illustration to large numbers of small domains, also of categories of the estimand, and of categories of the associated variable dimension, are straightforward. We note that where several associated variables are used, the combination of categories of these variables are strung out in a single dimension. The important point of this approach is that the estimation process is completely specified by the two data structures; the association and allocation structures. In summary, we require our estimation procedure:

(i) to conform to the current information in the allocation structure; and to

(ii) preserve somehow the earlier relationships present in the association structure without interfering with aim (i) above.

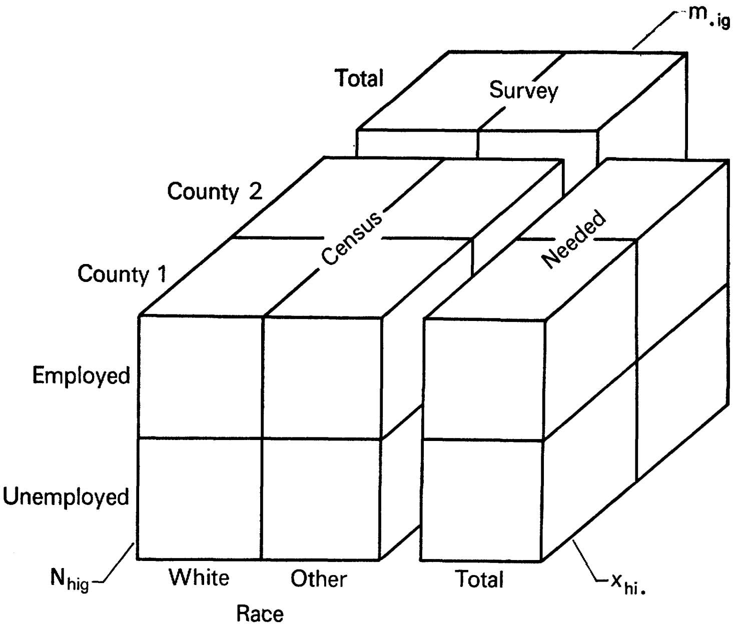

Figure 1. *A simple example of the data structure for small domain estimation.*

Then depending on what information is incorporated into the association and the allocation structures, the form of the resulting estimates that satisfy (i) and (ii) will differ. For this general method we now examine the principal specific techniques distinguished by the nature of the information incorporated into the association and into the allocation structures.

(a) $N = \{N_{hig}\}$ and $\boldsymbol{m} = \{m_{\cdot ig}\}$.

We present this first as the most important situation: information is full (h, i, g) in the association structure, but the small areas (domains) dimension h is unavailable from the current sample data in the allocation structure. We shall show the superiority of this full structure over the ordinary synthetic estimator, the BASE estimator of model (d), in Section 5.

Several criteria can be used to estimate $\boldsymbol{x}$ in this important case and they all lead to the same estimate. Probably the simplest one is a weighted least squares approach to minimize a weighted sum of squared differences between the original cell counts, N_{hig}, and the resulting modified cell counts, x_{hig}, subject to the marginal constraints specified by the allocation structure. That is, by standard Lagrangian analysis we minimize the function

$$F\,(x_{hig}) = \sum_{hig} w_{hig}^{-1}\,(x_{hig}-N_{hig})^2 - \lambda\left(\sum_{h} x_{hig} - m_{\cdot ig}\right),$$

where w_{hig} are some prespecified weights. If we choose $w_{hig} = N_{hig}$, then our estimator becomes

$$x_{hig} = \frac{N_{hig}}{N_{\cdot ig}}\,m_{\cdot ig}.$$

The resulting estimates for our variable of interest, i, in each small domain, h, are therefore

$$x_{hi} = \sum_{g} \frac{N_{hig}}{N_{\cdot ig}}\,m_{\cdot ig}. \tag{1}$$

The same estimator is obtained if we use maximum likelihood, also if we minimize a discriminant information criterion (see Purcell, 1979). These estimates are seen to be members of the general class of *segmented ratio estimators*, where the associated variable dimension defines the segments.

(b) $\boldsymbol{N} = \{N_{hig}\}$ and $\boldsymbol{m} = (\{m_{\cdot ig}\}, \{m_{h\cdot\cdot}\})$.

In addition to the information in (a), there are current estimates of totals for the local areas h. This addition can greatly improve the estimates (as we will see in Section 6). Current information on the sizes of the small domains may come from registers or from sample surveys, and may be on all areas or only on a sample of them. In this case, we could proceed by using any of the criteria mentioned in case (a), but the resulting Lagrangian expressions are not very tractable; obtaining an analytic solution is difficult. An alternative approach follows from an iterative procedure of Deming and Stephan (1940) for a similar adjustment problem. Their procedure is generally referred to as *iterative proportional fitting* (IPF), although other names are also used: iterative scaling, marginal raking and marginal scaling. The IPF solution only approximates the weighted least squares solution, but it simultaneously maximizes the likelihood equation of the multinomial distribution and minimizes the discriminant information (in the context of simple random sampling assumptions).

As we have already seen in (a), when there is only one set of margins defining the allocation structure, a simple solution exists which is just a segmented ratio estimator. The approach proposed here utilizes this fact by successively adjusting the cell frequencies, N_{hig}, to agree with the two sets of marginal constraints, $\{m_{\cdot ig}\}$ and $\{m_{h\cdot\cdot}\}$, through the use of (1) in an iterative cyclical fashion. Thus the procedure is as follows:

At the first step the starting values are set equal to the cell counts specifying the association structure. That is,

$$x_{hig}^{(0)} = N_{hig}.$$

These cell counts are then adjusted to the first set of marginal constraints, specified by the allocation structure, $\sum_h x_{hig} = m_{\cdot ig}$, then to the second set of constraints, $\sum_{ig} x_{hig} = m_{h\cdot\cdot}$, in an iterative fashion. Thus an iterative cycle consists of two steps, and at the kth iteration we have

$$_1x_{hig}^{(k)} = \frac{x_{hig}^{(k-1)}}{x_{\cdot ig}^{(k-1)}} m_{\cdot ig} \quad \text{and} \quad x_{hig}^{(k)} = \frac{_1x_{hig}^{(k)}}{_1x_{h\cdot\cdot}^{(k)}} m_{h\cdot\cdot},$$

where $_1x_{hig}^{(k)}$, are the estimates resulting from adjusting to the marginal constraints, $\{m_{\cdot ig}\}$ at the kth iteration. The resulting estimates, $x_{hig}^{(k)}$, are then used as inputs into the next cycle. The iteration is continued until some convergence criterion is satisfied following an iteration cycle. These estimators can be said to be members of the general class of *two-step raking ratio estimators.* Once again we sum the resulting estimates, x_{hig}, across the associated variable dimension to give our final *two-step segmented raking ratio estimates.*

(c) $\boldsymbol{N} = \{N_{hig}\}$ and $\boldsymbol{m} = (\{m_{\cdot ig}\}, \{m_{h\cdot g}\})$.

In this case, there is more additional current information, and the solution follows directly from the developments in (b). The only difference is that at the kth iteration the second step adjustment is given by

$$x_{hig}^{(k)} = \frac{_1x_{hig}^{(k)}}{_1x_{h\cdot g}^{(k)}} m_{h\cdot g}.$$

Again by summing the solution over the associated variable dimension we obtain two-step segmented raking ratio estimates.

(d) $\boldsymbol{N} = \{N_{h\cdot g}\}$ and $\boldsymbol{m} = \{m_{\cdot ig}\}$.

This represents a common situation when the full association structure $\boldsymbol{N} = \{N_{hig}\}$ is not available for the estimation process because the estimand data have not been collected at the earlier (census) time; for example, the use or non-use of contraceptives. Rather, the census data define an *incomplete association structure*, which is represented by the cross-tabulation of the associated variables by small domains at the previous census. Thus the information we have on the association structure is defined by $N = \{N_{h\cdot g}\}$. Because this structure lacks (or ignores) information for categories i of the estimand variable, a model must be used to define a *dummy association structure*. The simple and common procedure is to assume proportionality across the i categories and use

$$x_{hi} = \sum_{g} (N_{h\cdot g}/N_{\cdot\cdot g}) m_{\cdot ig}. \tag{2}$$

This is a special case of the segmented ratio estimator, which we term the basic synthetic estimator (BASE). We discussed this in 2.2, as well as another 'synthetic' estimator,

$$\bar{x}_{hi} = \sum_{g} (N_{h\cdot g}/N_{h\cdot\cdot}) \bar{m}_{\cdot ig}. \tag{3}$$

Both these estimators use only $N_{h\cdot g}$ instead of N_{hig}, which is useful when the N_{hig} are not available. But when these are available, the estimators (2) and (3) impose a rigid model needlessly and wastefully, as we show in the increased biases in Section 6; the alternative model (a) yields considerably reduced biases in our empirical study.

(e) $N = \{N_{h\cdot g}\}$ and $\boldsymbol{m} = (\{m_{\cdot ig}\}, \{m_{h\cdot\cdot}\})$.

(f) $N = \{N_{h\cdot g}\}$ and $\boldsymbol{m} = (\{m_{\cdot ig}\}, \{m_{h\cdot g}\})$.

The estimators in these cases follow on directly from cases (b) and (c), except that we utilize the dummy association structure, as in (d). Thus, the only difference is that our starting values are now defined by

$$x_{hig}^{(0)} = N_{h\cdot g}, \quad i = 1, \ldots, I.$$

This approach, as we mentioned before, can be easily modified to address the problem of estimating small domain non-frequency characteristics such as total income or expenditure. Essentially, this involves redefining the association and allocation structures in terms of the total value for the variable of interest in each cell, rather than cell counts (frequencies). For example, the association structure would be defined in case (a) as $N = \{Y_{hig}\}$, where the components Y_{hig} represent the total value for the ith category of the estimand in the hth small domain and gth category of the associated variable dimension. Usually, for non-frequency characteristics the estimand variable would only have one category. Per capita estimates can then be obtained by dividing the estimates of total for each small domain by the respective estimated (or known) population sizes, in the appropriate cells.

Each of the estimators derived above can be expressed in log-linear form as shown by Purcell (1979). This fact can then be utilized to show that all the proposed estimators preserve (carry over) all the interactions specified by the association structure except those that are respecified by the allocation structure. In other words, all the relationships between the variables that are incorporated into the association structure (past data) will hold in the current estimates, except those that are restated by the allocation structure (current data).

Consequently, as a function of this structuring of the small domain problem within a framework of categorical data analysis the resulting estimators all carry over certain interactions from the association structure. Thus all the estimators developed here can be said to belong to a class of *structure preserving estimators* (SPREE). The effects in the different cases can be compared in Table 1. In case (a) the one-way effects i and g and their two-way interaction ig come from the current (new) data ($m_{\cdot ig}$), but the others (h, hi, hg and hig) come from the earlier (census) data. Case (d), the BASE synthetic estimator, differs in that the hi and hig interaction are both forced to be zero in the association structure, and in the estimates.

The implementation of the SPREE estimation procedure is a straightforward process, and Purcell (1979) documents a Fortran program that carries out this estimation. Basically, it requires only the input of the association and allocation structures.

Table 1

Structure preserving properties of different SPREE estimates

	Effects						
	One-way			Two-way			Three-way
Estimator	*h*	*i*	*g*	*hi*	*hg*	*ig*	*hig*
(a)	*c*	*n*	*n*	*c*	*c*	*n*	*c*
(b)	*n*	*n*	*n*	*c*	*c*	*n*	*c*
(c)	*n*	*n*	*n*	*c*	*n*	*n*	*c*
(d)	*c*	*n*	*n*	*o*	*c*	*n*	*o*
(e)	*n*	*n*	*n*	*o*	*c*	*n*	*o*
(f)	*n*	*n*	*n*	*o*	*n*	*n*	*o*

Note: n = new; c = census; o = zero.

5 Comparison of different SPREE estimators

The variances of the SPREE estimates depend mostly on the variances of the marginal constraints specified by the allocation structure, because the association structure is derived from past census data. Thus, the variances of these estimates will, in most practical situations, be small compared to the biases which are less controllable, and will have a greater influence on the efficiency of the different SPREE estimators.

The SPREE estimates tend to be biased estimates to the degree that the underlying association structure, which gets imposed on the estimates, does not correctly represent the true structure present in the current population. The bias question is obviously closely related both to the choice of marginal constraints imposed through the allocation structure, and to the form of the association structure used. Thus, it can be expected that the different SPREE estimates developed above will have considerably different magnitudes of bias.

Purcell (1979) has carried out an empirical investigation into the bias of the different SPREE estimates developed in the previous section. We present a brief summary of its major findings and of their implications for efficient small domain estimation. In the study, vital statistics and census data were used to obtain 'synthetic' estimates of mortality due to each of four different causes (the estimands i) and for each state (domains h) of the United States. The associated variable dimension, g, was defined as 36 convenient age–sex–race groups. Estimates and biases were calculated for five individual years ranging over a full 10-year postcensal period, 1960 through 1970, so that, as the association structure becomes increasingly out of date, the performance of the different estimators could be evaluated.

Due to the nature of the data sources used, the resulting SPREE estimates have zero sampling variability and no conceptual bias. The only source of error in the estimates is the *SPREE estimate bias*, which can be measured in terms of the *percentage absolute relative difference* (%ARD) defined in this case as

$$\%\text{ARD} = \frac{|x_{hi\cdot} - X_{hi\cdot}|}{X_{hi\cdot}} \times 100,$$

where $x_{hi\cdot}$ and $X_{hi\cdot}$ are the SPREE estimate and the true number, respectively, of total deaths by specific cause i, for the hth state. Several different SPREE estimators were evaluated in this study and they are specified in Table 2.

We use medians of the state %ARDs here (Table 3) because of the skewness of the distributions, but the means yield very similar results (Purcell, 1979). It is clear that the accuracy (in terms of the method bias) of the different SPREE estimates varies considerably. Differences are also found within each estimator, both between the 4 causes of death studied, and with the length of the postcensal period. However, only the level of the bias is different, and the

Table 2

Classification of the different SPREE estimates evaluated

	Allocation Structure	
Association Structure	$(\{m_{\cdot ig}\}, \{m_{h\cdot\cdot}\})$	$\{m_{\cdot ig}\}$
N_{hig}	*b*	*a*
$N_{h\cdot g}$	*e*	*d*

Table 3

Medians of the percentage absolute relative differences of different SPREE estimates

		Cause of Death			
Estimator	Year	Malignant Neoplasms	Major CVR Diseases	Suicides	Total Other
(b)	1960	—	—	—	—
	1961	1.853	0.732	6.488	1.386
	1964	2.221	1.028	8.642	2.199
	1967	3.222	1.197	6.317	3.355
	1970	2.752	2.218	8.518	3.854
(a)	1960	—	—	—	—
	1961	1.975	1.470	5.556	1.923
	1964	3.502	1.981	8.983	3.275
	1967	5.581	3.471	7.761	4.892
	1970	8.183	4.716	13.415	6.650
(e)	1960	6.310	3.362	12.162	7.357
	1961	5.620	2.963	12.941	7.163
	1964	5.783	2.513	14.444	6.152
	1967	5.588	2.215	15.238	5.033
	1970	5.606	2.739	11.943	4.867
(d)	1960	9.255	8.417	11.862	6.700
	1961	9.171	7.506	18.519	5.717
	1964	10.053	7.116	15.306	7.191
	1967	10.070	7.916	17.333	6.457
	1970	8.503	11.111	19.417	7.736

pattern of the differences is basically the same for each of the four causes of death. Fig. 2, which illustrates these differences for deaths due to malignant neoplasms is typical of the pattern for the other causes of death.

The most striking feature of the analysis of the different SPREE estimates is the strong performance of the estimates based on the full association structure ((a) and (b)). While the (a) estimates (and to a lesser degree the (b) estimates) did show a tendency to deteriorate over time, they were clearly superior to the corresponding estimates based on the incomplete association ((d) and (e)). The performance of (b) compared to (a) points to the importance of incorporating through the allocation structure the maximum available current information into the estimation process. The association structure, however, must be accurate as the SPREE estimation procedure is very sensitive to approximations (and bias) in the allocation structure (Purcell, 1979).

The overall performance of the estimates based on the incomplete association structure ((d) and (e)) is inferior. The performance of the commonly used BASE estimator (d), in what can be regarded as a well behaved data set, raises some concern over its wide use. Its performance, however, is seen to be somewhat strengthened by the use of additional accurate current information as in the (e) estimates.

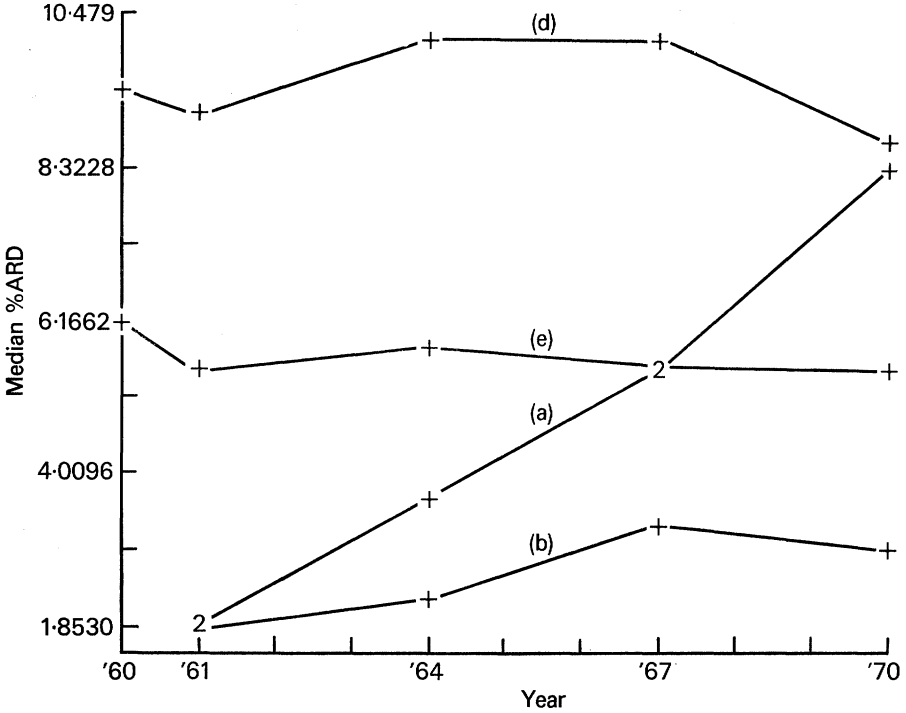

Figure 2. *Median %ARD of different SPREE state estimates of the deaths due to malignant neoplasms by year.*

Clearly, estimates with low bias can be obtained if accurate association and allocation structures are available on which to base the estimation. But when we seek to input less information and approximate these structures, either for simplicity or due to data availabilities, the performance falls away.

6 Some modifications of the SPREE approach

The use of the dummy association structure, when the variable of the interest has not been collected in the previous census, is in effect equivalent to using an unsaturated log-linear model to approximate the full association structure; various interaction terms are forced to be zero. Such reduced association structures may, however, be utilized for several other reasons:

1. The full association structure may not be available because the required data were not collected in the census (e.g., in models (d) and (e) in Section 4). Or, even if collected, those data from the census were simply not published at the required levels. For example, the three-way cross-classification, $h \times i \times g$, may not be available from census tabulations, but the three two-way cross-classifications, $h \times i$, $h \times g$ and $i \times g$ may be published. And it may be too expensive or difficult to go beyond published tabulations. Special tabulations from census tapes can be extremely costly. But, from the available partially overlapping lower order cross-classifications a 'dummy' association structure can be built by raking methods. In this, however, those higher order interactions that are not specified by the lower order cross-classifications are set to zero. The appropriateness of this procedure depends on whether these interactions are actually negligible. Otherwise bias results.
2. When the number of cells in the association structure is very large, it can frequently occur that many of the cell frequencies are zero or very small. This problem of 'thin' data is

often handled by collapsing across adjacent cells. Alternatively it could be handled by using some reduced association structure; this again usually involves setting certain interactions to zero.

3. The SPREE procedure, discussed in Section 4, preserves all the interactions specified in the association structure except those redefined by the allocation structure. However, by assuming that all 'preserved' interactions are stable over the postcensal period we are introducing a significant bias, whenever unstable subsets are present. One possible approach, where we suspect 'unstable' interactions, is to preserve only the interactions regarded as stable. This can be done by fitting an unsaturated log-linear model to the observed association structure, from which adjusted starting values are obtained for SPREE estimation, which reflect the required reduced association structure. However, a difficulty with this is that the 'unstable' interactions are set equal to zero, and it is unclear whether this leads to more harm than leaving in the 'wrong' interactions, as we do when basing the association structure on the complete saturated model.
4. It should be possible, instead of either accepting obsolete interactions or setting them to zero, to modify them according to some theoretical/empirical model. But this is for future research.

Case (1) is of prime interest concerning the question of simplified data input requirements for efficient small domain estimation, and the mortality data were used to test it (Purcell, 1979). The full association structure was ignored and we assumed the past data only provided us with the three two-way cross-classifications of the variables. By raking of these margins we are effectively fitting the reduced association structure that specifies that the three-way interaction is zero. On this reduced association structure, we can then impose the two different allocation structures used in the evaluation in Section 5. The estimates based on the reduced association structure are denoted by

Estimator	*Allocation structure*
(g)	$\{m_{\cdot ig}\}$ and $\{m_{h\cdot\cdot}\}$
(h)	$\{m_{\cdot ig}\}$

To investigate the equivalence of the corresponding estimates based on the full and reduced structures, the estimates were compared directly to each other. The main results of this comparison are presented in Table 4, where the median state %ARDs between the full and reduced association structure based estimates are summarized. The %ARDs calculated are given by

$$\%\text{ARD (1)} = \frac{g-b}{b} \times 100$$

and

$$\%\text{ARD (2)} = \frac{h-a}{a} \times 100.$$

From Table 4, we see that the full and reduced association structure based SPREE estimates are essentially equivalent, and it was only the suicide estimates that show any real differences (the maximum median %ARD was still only 2.82 per cent). However, the differences that did occur showed a slight tendency to increase uniformly over time, for all causes of death and for all comparisons. While the reasons for this are not entirely obvious, we suspect it is mainly a reflection of the increase in the variability of the estimates themselves, over time.

While these results are specific to these data, they do lend support to the contention that, in other small domain data, higher order interactions can be ignored and set to zero. Clearly, a wider scope of data needs to be examined, as such a result is of importance from the point of view of reduced data input requirements for efficient SPREE estimation.

Table 4

Medians of the percentage absolute relative differences between full and reduced structured SPREE estimates

		Cause of Death			
%*ARD*	Year	Malignant Neoplasms	Major CVR Diseases	Suicides	Total Other
%*ARD*(1)	1961	0.0406	0.0142	0.5450	0.0398
	1964	0.0637	0.0350	1.1594	0.0806
	1967	0.1081	0.0623	1.9084	0.1503
	1970	0.1145	0.0897	2.7027	0.1807
%*ARD*(2)	1961	0.0456	0.0204	0.5063	0.0381
	1964	0.0737	0.0334	1.1765	0.0517
	1967	0.1546	0.0810	1.9553	0.1282
	1970	0.1945	0.1344	2.8210	0.1514

7 Some concluding remarks

It should be stressed that we view the categorical data analysis approach as a valuable though not exclusive method for efficient small domain estimation. We recognize the need for composite procedures combining the SPREE estimation approach with other methods, in order to fully utilize the different strengths of the various estimation procedures. This can best be facilitated by formulating a unified framework, incorporating considerations of cost effectiveness for evaluating the different estimation strategies.

The importance of such a unified framework for future small domain estimation developments cannot be stressed enough. Among other things, this framework should incorporate resource/methods considerations relating to the use – either singularly or in combination – of both direct and indirect estimation for small domains. Also important is an assessment of the cost-benefits of straightforward versus complex computational algorithms for both the small domain estimates and their subsequent evaluation.

Direct estimation can be strengthened with more current censuses (quinquennial?) or microcensuses (annual?), or with larger sample surveys, or with more complete and better registers. But various and legitimate constraints of resources are bound to intervene. Demands for timely, detailed and complete data are bound to run ahead of the supply of data. Thus there will be considerable scope and need for strengthening indirect procedures for small domain estimation; such as regression and especially SPREE methods. The framework provided here should be useful not only for better analyses, but also for pointing out the kind of data that would be most useful for better estimates. In general, the central issue involved in using indirect methods is the appropriateness of the underlying models, arising either implicitly through the basic assumptions on which the techniques incorporating this information are based, or directly through our assumptions about the data structures. Therefore, the quality of indirect small domain estimates depends largely on being able to identify associated variables that are highly related to the variable of interest, and on understanding the data structures (relationships) existing between the variables.

Such a total design approach – incorporating considerations of survey design, data analysis and data management considerations simultaneously – will enable us to better meet the increasing small domain data requirements of our societies. Within this framework, we anticipate further developments that will expand theory and applications to new problems and to new variables. The ultimate aim is more efficient procedures for combining the strengths of the diverse sources of data – samples, censuses and registers – to construct needed *synthetic small domain data bases*. Estimates derived from these data bases should be simultaneously more detailed, timely, and accurate than those currently available.

Finally *forecasting* for local areas and other small domains poses a clear and present challenge to our methods. It appears as an extension of postcensal methods which also go beyond the last census. But, our postcensal methods are based on current marginal data which are missing for forecasts. However, it should be possible to base forecasts on models which forecast a continuation of the movement of current and past interactions (relationships) into national projections of total populations.

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 No. 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.

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. *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.

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.

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. (1979). Samples and censuses. *International Statistical Review*, **47**, no. 2.

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 Kish, L. (1979). Estimation for small domains. *Biometrics*, **35**, 365–384.

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.

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

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.

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 Use 6, No. 4*, 18–27.

Thompson, J.R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, **63**, 113–122.

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 (1975*a*). 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 (1975*b*). *The Methods of 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.