Journal of Official Statistics, Vol. 15, No. 1, 1999, pp. 1–24

Organization of Small Area Estimators Using a Generalized Linear Regression Framework

David A. Marker [1]

In this article existing small area estimators are described, including Bayesian ones that have been proposed. Ghosh and Rao (1994) provide an excellent description of many of the techniques found in the current literature, pointing out the importance of the level of aggregation at which the models are developed. In this article a literature review is conducted for the estimators. The estimators are then organized from a general linear regression perspective, summarizing and showing where certain methods can be viewed as minor variations or generalizations of others. This includes a derivation of the conditions under which it is possible to view synthetic estimation as a form of regression. The goal of this article is to pull together the wide range of approaches that have been used for small area estimation. From this review a clearer understanding of the present techniques and their interrelationships is apparent.

Key words: Bayes; synthetic estimation; components of variance regression.

1. Introduction

For the last 25 years the special problems of deriving estimates for small areas or domains (subsets of the entire population) from sample surveys have received increasing attention within the survey sampling literature. Many attempts to derive such estimators have been either ad hoc approaches for specific problems, assuming specific models for the data, or attempts to apply large-sample sampling theory to problems of small samples. In this article existing small area estimators are described, including Bayesian ones that have been proposed. The goal is to develop a more coherent understanding of the problem and methods with which such estimation can be undertaken.

The typical survey sampling approach to estimation is to design the survey to produce design-consistent estimates, that is, estimates whose expectations over the set of all possible samples are consistent for the unknown parameters. Kish (1987) points out that not all domains for which estimates are desired are ''major domains,'' that is, domains of the population for which design-based estimates of acceptable precision can be produced. Frequently estimates for smaller ''minor domains'' are desired but the standard design-based estimates are too unstable.

Minor domains of interest depend on the subject matter. In many demographic and health surveys, minor domains of interest include geography, and combinations of race,

[1] Westat Inc., 1650 Research Blvd., Rockville, Maryland 20850, U.S.A. E-mail: MARKER1:@westat.com.

Acknowledgments: I would like to acknowledge the guidance of my Ph.D. dissertation committee Dick Cornell, Jim Lepkowski, and Bob Groves of the University of Michigan and Graham Kalton of Westat. The statistical group at Westat, and in particular David Morganstein and Mary Lou Pieranunzi, were very helpful throughout this effort.

age, sex, and ethnicity. Examples of geographical small area estimates (derived from a national survey) that are frequently discussed are state-level health indicators, inter-censal population and housing estimates for states and counties, and poverty indicators for school districts. This article examines estimators only for geographical domains, although some of these estimators may also be appropriate for other types of domains.

When the domains of interest are known in advance it may be possible to design the sample to ensure that sufficient sample sizes are available for each domain to allow for precise design-unbiased estimates (Singh, Gambino, Mantel 1994). Auxiliary variables may be used to improve the precision of design-unbiased estimates. The traditional survey sampling ratio, regression, and post-stratified small area estimators are all methods for incorporating auxiliary data into the estimation procedure. Särndal and Hidiroglou (1989), Cassel, Kristiansson, Råbäck, and Wahlström (1987), and Ghangurde and Gray (1978) give examples of using such estimators, the first two modifying traditional regression estimators based on the set of observed residuals and the third using regression to adjust unbiased direct estimates. However, using auxiliary data in conjunction with the observed data in a non model-dependent manner may not be sufficient to provide precise estimates. Budgets may prevent one from allocating enough sample to all domains. Also, additional domains are often specified after the survey is designed. It is therefore common to encounter domains of interest for which the sample data are insufficient to provide precise design-unbiased estimates. It is then necessary to ‘‘borrow strength’’ from the data observed from other domains or time periods using some form of model-dependent estimator to improve the stability of the estimates.

Survey sampling practitioners, empirical Bayesians, and Bayesians frequently disagree on the philosophical basis for estimation, but small area estimation is one area where there is a consensus on the need for model-dependent estimation. This article reviews the different small area estimators developed by these three groups of statisticians (survey sampling practitioners, empirical Bayesians, and Bayesians) and explains the interrelationships between the different estimators.

2. Interrelationships Among Small Area Estimation Techniques

Researchers and practitioners with different statistical philosophies have developed different procedures for the derivation of small area estimates and measures of their accuracy. Ghosh and Rao (1994) provide an excellent description of many of the techniques found in the current literature, pointing out the importance of the level of aggregation at which the models are developed. In this article the estimators are organized from a general linear regression perspective, summarizing and showing where certain methods can be viewed as minor variations or generalizations of others. It is hoped that this will lead to a clearer understanding of the present techniques and their interrelationships.

The next subsection contains a review of estimators developed using the traditional survey sampling approach. Later subsections review empirical Bayesian and hierarchical Bayesian approaches. In Section 3 the interrelationships between the estimators are summarized.

2.1. *Traditional survey sampling approach to small area estimation*

It is appealing to consider sample designs such that direct estimates of adequate precision can be obtained for all domains of analytic interest from the sample data. The properties of direct estimators (Schaible 1993a) are well known and free of any dependence on models. Unfortunately, there is frequently a need for estimates for domains for which the direct estimators are subject to unacceptably large variances, if direct estimates can be obtained at all. Survey statisticians have proposed a number of small area estimators in an attempt to overcome this limitation.

This section will describe how these estimators can be placed in a general linear regression framework of the form $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + \epsilon$ where Y is a continuous variable, the X variables are (assumed) known predictor variables, and the B's are regression coefficients. The relationship of the estimators to regression is developed here to explain their commonalities and to give guidance on how the methods can be generalized. Various small area estimators make assumptions regarding whether the predictor variables, X, are continuous, ratios, or indicator variables, and also about the form of the error structure of Y.

Nine different survey sampling approaches are reviewed here. Seven of these are described by Purcell and Kish (1979): symptomatic accounting techniques (SAT), regression-symptomatic (also known as symptomatic regression or ratio-correlation) procedures, sample regression, synthetic estimation, the base unit method, synthetic regression, and structure preserving estimation (SPREE). Two others – the vital rates method and components of variance regression – have been added to their enumeration. Apart from the SAT and SPREE, all of these methods can be formulated in the regression framework. Only SAT is nonstochastic; it will be described briefly before discussing the other procedures. The SAT and vital rates methods are concerned exclusively with estimation of vital statistics or the population at the small area level, whereas the remaining methods are of more general application. Therefore, we discuss the SAT and vital rates methods before proceeding to the remaining methods.

2.1.1. Symptomatic accounting techniques

The SAT simply updates old census figures by adding or subtracting values of variables directly related to (symptomatic of) the change of interest. For example, in estimating population sizes for small areas symptomatic variables are births, deaths, immigration, and migration. If all factors influencing the variable of interest could be independently listed and measured, there would be nothing else needed. The updated population for an area could be simply estimated as: (previous population) + (registered births) − (registered deaths) + (immigration) − (emigration).

Unfortunately, births are not always easily assigned to small areas, some births may go unregistered, and migration is difficult to measure. Applications of the SAT assume this model to work without error. Since this small area technique is nonstochastic, it will not be considered further.

2.1.2. Vital rates

The SAT can be generalized (Bogue 1950) using the rates of change in births and deaths rather than the counts of births and deaths. Implicit in this vital rates method is the assumption that the ratio of the birth (or death) rate for a given small area to the birth (or death) rate for a larger region remains constant from one census to the next. (Birth rate is defined as the number of births divided by the size of the population.) If these rates have been stable or falling since the last census, then this assumption of a constant ratio of rates may be sound. If, however, the rates are rising, then this assumption of the constant ratio of rates assumption is unlikely to hold in practice.

To demonstrate how the vital rates method is a regression technique, consider for example the problem of estimating the population for small area i within state s, with the small areas a mutually exclusive and exhaustive partition of s. The number of births, B_{it}, is available from hospital or local records. The population of small area i is estimated from its birth rate, BR_{it}, which itself must be estimated. The vital rates technique assumption of a constant ratio of birth rates over time (for all small areas i in state s) is given by $(BR_{it}/BR_{st}) = (BR_{i(t-1)}/BR_{s(t-1)})$, where BR_{it} is the birth rate in small area i from time period $t-1$ to t, and BR_{st} is the birth rate in state s (which contains small area i) from time period $t-1$ to t.

It is assumed that the state-level rates $BR_{s(t-1)}$ and BR_{st} are available from official sources, as is the small area i birth rate from the previous period $BR_{i(t-1)}$. Thus, the current birth rate in small area i is, under the constant ratio assumption, given by

$$B\hat{R}_{it} = \frac{BR_{st}}{BR_{s(t-1)}} \cdot BR_{i(t-1)} \tag{1}$$

The vital rates estimate of the population total for small area i at time t is $Pop_{it} = B_{it}/BR_{it}$, where B_{it} is the number of births in small area i from time period $t-1$ to t.

The vital rates technique assumes that no knowledge is gained about state s or its small areas by examining other states; therefore, it is appropriate to develop a separate univariate regression model for each state. One possible univariate regression model for each state is a model without an intercept term:

$$Y_i = BX_i + e_i \quad E(e_i) = 0 \quad V(e_i) = \sigma^2 X_i/W_i \tag{2}$$

where $Y_i = BR_{it}$; $X_i = BR_{i(t-1)}$; $W_i = (N_{i(t-1)})/(N_{s(t-1)})$; $N_{i(t-1)}$ is the population in small area i at time $t-1$; and $N_{s(t-1)}$ is the population in state s at time $t-1$. W_i, the proportion of the state's population in small area i, is assumed not to have changed from time $t-1$ to t. The variance assumption $V(e_i) = \sigma^2 X_i/W_i$ implies small areas with large birth rates will be most variable, as will those small areas with smaller populations, N_i. The weighted least squares estimate of $\hat{B}$ is:

$$\hat{B} = \frac{\sum\limits_{i \epsilon s} Y_i X_i/(X_i/W_i)}{\sum\limits_{i \epsilon s} X_i^2/(X_i/W_i)} = \frac{\sum\limits_{i \epsilon s} W_i Y_i}{\sum\limits_{i \epsilon s} W_i X_i} = \frac{BR_{st}}{BR_{s(t-1)}} \tag{3}$$

It follows from Equation (2) that

$$B\hat{R}_{it} = \hat{B}X_i = \frac{BR_{st}}{BR_{s(t-1)}} BR_{i(t-1)} \tag{4}$$

This is equivalent to the vital rates assumption found in Equation (1). Thus, under the model in (2), the vital rates technique is simply the weighted least squares solution to univariate regression.

It is worth noting that if the assumption that the W_i are stable across time holds, the population in small area i could alternatively be estimated by $N_{it} = W_i N_{st}$, without using the model in (2). If the W_i vary across time Equation (3) can still be used but it will yield a weighted average of the BR_{it} that is not necessarily equal to BR_{st}.

2.1.3. Symptomatic regression

Symptomatic regression uses data from two past censuses to determine for a set of small areas the relationship between changes in the variable of interest and changes in a set of symptomatic variables. That relationship is then used to predict the change in the variable of interest for each small area from the most recent census to the current time from observed changes in the symptomatic variables during the same period.

The symptomatic regression procedure takes the ratio for each symptomatic variable of the most recent census to the preceding census as the independent variables for each small area. The corresponding ratio for the variable of interest is then regressed on the symptomatic variables. The same ratios for the symptomatic variables are derived using their present values and the most recent census data to obtain the updated values of these variables. Combining these current values of the symptomatic variables with the previously derived regression coefficients provides an estimate of the change since the last census in the dependent variable. By multiplying this number by the dependent variable's value at the time of that census the symptomatic regression estimate for the present period is determined.

The multiple symptomatic regression method uses the model $\mathbf{Y} = X\mathbf{B} + \epsilon, \epsilon \sim N(\mathbf{O}, \Sigma)$; where $\mathbf{Y}$ is an $(I \times 1)$ column vector $[Y_1, Y_2, \ldots, Y_I]'$; $Y_i, i = 1, \ldots, I$, are the ratios for the dependent variable in each small area i of the most recent census to the preceding census; X is an $(I \times p)$ design matrix of $X_{ir}, i = 1, \ldots, I$ and $r = 1, \ldots, p$, the ratios of the most recent census values to the preceding census values for area i for the p symptomatic variables (e.g., number of births or income tax forms filed); and $\Sigma = \text{diag}\{\sigma^2\}$, an $I \times I$ matrix. The least squares estimate for $\mathbf{B}$ is then $\hat{\mathbf{B}} = (X'X)^{-1}X'\mathbf{Y}$.

The small area symptomatic regression estimation procedure uses the estimate $\hat{Y}_i^* = X_i^*\hat{\mathbf{B}}$, where X_i^* are the p symptomatic variable ratios of their present value to their value at the time of the most recent census for small area i. The estimate of change $\hat{Y}_i^*$ is then multiplied by the most recent census value for the small area to give the symptomatic regression estimate of the present total for small area i. As with the vital rates method, symptomatic regression is thus a form of multiple regression on a set of symptomatic variables.

Purcell and Kish (1979) noted that this method is based on the assumption that the statistical relationship between the independent and dependent variables in the last intercensal period will remain the same in the current postcensal period. In addition to depending on the stability of this relationship over time, the symptomatic regression procedure's accuracy relies on the strength of the multiple correlation of the symptomatic variables with the dependent variable.

Bogue's vital rates method can be generalized into the multivariate framework (Schmitt and Crosetti 1954) by choosing among a variety of symptomatic variables based on their historical correlations with the dependent variable. Current versions of the vital rates method would use multiple correlations (multiple regressions) rather than univariate associations. Symptomatic regression has been applied to produce estimates of population

(Bogue and Duncan 1959), retail trade (Woodruff 1966), and single-family housing permits (Otelsberg 1981). Pursell (1970) and Martin and Serow (1978) examined the utility of including dummy variables in the symptomatic regression as a form of pseudo-stratification.

Some researchers have also used symptomatic regression in combination with other small area techniques. For example, the U.S. Bureau of the Census has averaged three different methods, multiple symptomatic regression, administrative records, and component method II (U.S. Department of Commerce 1974 and 1980), to derive its population and per capita income estimates for states, counties, and sub-county areas. Four symptomatic variables are used in the symptomatic regression: school enrollments, number of Federal income tax returns, car registration, and size of work force. The administrative records and component method II methods use symptomatic accounting techniques to track births, deaths, and the elderly, and the vital rates methodology to track migration at the sub-county level. Just as multiple symptomatic regression and vital rates have been shown to be a form of regression, so too is the U.S. Bureau of the Census's method.

2.1.4. Sample regression

One of the major drawbacks of the symptomatic regression technique is that it assumes a constant relationship between the independent and dependent variables reaching back to the census preceding the most recent one. The sample regression method introduced by Ericksen (1973), and based on earlier work by Hansen, Hurwitz, and Madow (1953), instead makes use of sample data on the dependent variable for a sample of small areas. The sample regression method can use any variables that are available for all small areas as explanatory variables, but in all of its small area applications the changes in symptomatic variables since the most recent census have been used. As with symptomatic regression, Ericksen's method requires a linear relationship among the variables.

The sample regression method follows a similar procedure to symptomatic regression but has a less demanding set of assumptions. Rather than estimating the regression coefficients from ratios of symptomatic variables from one census to the next, sample regression uses the relationship between aggregate-level sample estimates and explanatory variables for a sample of small areas. The coefficients are computed from a regression that is only computed for those small areas with available sample data, but these coefficients are then applied to all small areas. Traditionally, the sample regression estimator has used ratios of current values to the most recent census value for both the dependent and explanatory variables. It is this historical version of the sample regression estimator that is described below.

Suppose that sample data are available for only n of the I small areas. The sample regression method can be written as $\mathbf{Y} = X\mathbf{B} + \epsilon$, $\epsilon \sim N(O, \Sigma)$; where $\mathbf{Y}$ is an $(n \times 1)$ column vector of $y_i, i = 1 \ldots, n$, the ratios of the most recent small area sample estimate in area i to the previous census value for small area i; X is an $(n \times p)$ design matrix of $X_{ir}, i = 1, \ldots, n$ and $r = 1, \ldots, p$, the ratios of present values to previous census values in area i for the p symptomatic variables (e.g., number of births or income tax forms filed); and $\Sigma = \sigma^2 I + T$, where I is an $n \times n$ identity matrix and $T = \text{diag}\{\tau_i\}$ is a diagonal matrix of sampling errors for each small area i. The weighted least squares estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$ is used with the X_{ir} to compute the estimates for all of the I desired small areas in the same manner as with symptomatic regression, with $\hat{\mathbf{B}} = (X'\Sigma^{-1}X)^{-1}X' \sum^{-1} \mathbf{Y}$.

Sample regression no longer requires consistency of the relationship between the X and the Y over more than one census period but assumes that the relationship holds across all areas i since the last census. The trade-off is the additional variance from the two-stage sampling procedure, first choosing a sample of small areas and then sampling within those areas. The sample regression method also depends on the representativeness of the two-stage sampling process, choosing a sample of areas i and then sampling within those areas. If the sample of areas (or the samples within the areas) is not representative (with respect to the independent variables) of the full range of areas, extrapolating the regression coefficients to other areas may produce poor estimates. The less-restrictive set of assumptions shows the sample regression method to be a generalization of the regression approach to small area estimation.

The sample regression method has been applied to the undercount in the 1980 U.S. decennial census (Ericksen and Kadane 1987). Successful extrapolation to nonsampled small areas is dependent on the sampled areas being representative of the range of values taken by the independent variables used in the regression equation. If it is anticipated that small area estimates will be developed using this approach, the sample design for the survey should include this consideration when first-stage stratification variables are determined.

This approach can be improved on for small areas that were included in the sample (Ghosh and Rao 1994). For these areas a weighted combination of the direct small area estimator and the sample regression estimator can be produced (see the discussion in Section 2.2). For the many small areas without sample data this does not provide any improvement over the sample regression estimator.

2.1.5. Components of variance regression

Vital rates and symptomatic regression do not rely on any sample data to produce small area estimates, instead relying on stable relationships across time. Sample regression uses sample data at the small area level for a sample of small areas. Unlike the earlier regression models that are modeled at the small area level, the components of variance regression model is at the element level.

In the components of variance regression method, the error term has two components, one random and the other small area specific (Fuller and Harter 1987). The model is $\mathbf{Y} = X\mathbf{B} + \epsilon$, where each element of ϵ, ϵ_{ik}, can be divided into an area-level error term, v_i, and an element-level error term, u_{ik}, that is, $\epsilon_{ik} = v_i + u_{ik}$, where $i = 1, \ldots, n$; $k = 1, \ldots, n_i$; and n_i is the number of individual elements (e.g., people, establishments) sampled in small area i. $\mathbf{Y}$ is a $(\Sigma n_i \times 1)$ column vector of y_{ik} and X is a $(\Sigma n_i \times p)$ design matrix (p is the number of independent variables); $\boldsymbol{V}$ is a $(\Sigma n_i \times 1)$ column vector with the first n_1 rows v_1, the next n_2 rows v_2, etc., $\boldsymbol{U}$ is a $(\Sigma n_i \times 1)$ column vector of u_{ik}, $\boldsymbol{V} \sim N(\mathbf{O}, \Sigma_{vv})$, and $\boldsymbol{U} \sim N(\mathbf{O}, \Sigma_{uu})$.

The least squares regression estimates, $\hat{\mathbf{B}} = (X'X)^{-1}X'\mathbf{Y}$, are computed for $\mathbf{B}$. The average residual for small area $i, \bar{\epsilon}_i$, is then computed for each small area. Given the assumed error structure, $E(v_i|\bar{\epsilon}_i) = \bar{\epsilon}_i G_i$, where $G_i = (\sum_{vv} + n_i^{-1}\Sigma_{uu})^{-1}\Sigma_{vv}$.

Thus it is possible to produce estimates of v_i for each small area. This allows each small area to have an area-specific component, while the regression equation provides an overall estimate across all small areas. It is important to observe that v_i cannot be estimated for small areas i without any sample data. In such cases v_i is estimated as zero.

The components of variance approach can be expanded to include multivariate regression (Fay 1987) or to allow each sub-area (e.g., counties), in addition to each small area (e.g., states), to have a separate means by splitting v_i into two terms, one for the sub-area within the small area and one for the small area itself (Pfeffermann and Barnard 1991). The Pfeffermann and Barnard model defines the error term as $\epsilon_{ikl} = v_i + u_{ik} + w_{ikl}$, where $l = 1, \ldots, n_{ik}$, the number of individual sampled elements in the kth sub-area within small area i; and $\boldsymbol{V} \sim N(\mathbf{O}, \Sigma_{vv})$, $\boldsymbol{U} \sim N(\mathbf{O}, \Sigma_{uu})$, and $\boldsymbol{W} \sim N(\mathbf{O}, \Sigma_{ww})$ are each $(\Sigma_i \Sigma_k n_{ik} \times 1)$ column vectors.

2.1.6. Synthetic estimation

Synthetic estimation is different from the symptomatic or sample regression methods. Synthetic estimation in its simplest form first obtains national or regional subgroup estimates, say by age-race-sex subgroups, and then derives a small area estimate by taking the appropriate weighted average of these national subgroup estimates where the weights reflect the age-race-sex composition of the small area. The accuracy of this method depends on how similar each population subgroup's national/regional average is to its small area average and also on the accuracy of the weights.

Before presenting synthetic estimation in a regression framework, consider a standard formulation. The synthetic estimator of the average of characteristic Y for small area i is:

$$\hat{\bar{Y}}_{i..} = \left(\sum_j N_{ij} \bar{y}_{.j.} \right) / N_{i.} \tag{5}$$

where N_{ij} denotes the size of the population in small area $i, i = 1, \ldots, I$, and subgroup $j, j = 1, \ldots J$; $N_{i.}$ is the total population in small area i, and $\bar{y}_{.j.}$ is the sample average of Y for subgroup j, across all small areas.

Synthetic estimates were first proposed in 1968 by the National Center for Health Statistics to obtain state estimates for disability rates from the National Health Interview Survey. Bogue and Duncan (1959) foreshadowed the use of synthetic estimates when they suggested generalizing the vital rates version of symptomatic accounting techniques by estimating each population subgroup separately using the symptomatic variable most highly correlated with that group, and then combining the subgroups to derive an estimate of the small area. This can be thought of as a synthetic estimate based on historical rather than on sample data.

Using a superpopulation approach (see Section 2.2), Holt, Smith, and Tomberlin (1979) examined whether synthetic estimation can be placed in the general linear model framework as $Y_{ijk} = B_j + e_{ijk}$ where Y_{ijk} denotes the kth sampled individual of small area i and subgroup j, B_j is the mean value for subgroup j, and e_{ijk} is the error term distributed $N(O, \sigma^2)$ for $i = 1, \ldots, I$; $j = 1, \ldots, J$; and $k = 1, \ldots, N_{ij}$. This is a one-way analysis of variance model that assumes, as in synthetic estimation, that the expected value of all individual units belonging to a subgroup j will be equal, regardless of the small area they come from. The least squares and best linear unbiased estimate (BLUE) of B_j is $\hat{\mathrm{B}}_j = \bar{y}_{.j.} = \Sigma_i \Sigma_{k \in s} y_{ijk} / \Sigma_k n_{ij}$ where $k \epsilon s$ denotes those individual units in the sample and

n_{ij} denotes the number of sampled elements in small area i and subgroup j. The best linear unbiased predictor (BLUP) of the average in small area i is given by

$$\hat{\bar{Y}}_{i..} = \left(\sum_j n_{ij}(\bar{y}_{ij.} - \bar{y}_{.j.}) + \sum_j N_{ij} y_{.j.} \right) / N_{i.} \tag{6}$$

The synthetic estimate is the last term in Equation (6). The BLUP is thus not the synthetic estimate unless there are no sampled data from small area i (in which case the first term in Equation (6) is equal to zero). Using the superpopulation approach, therefore, does not lead to the classical form of the synthetic estimator except in an extreme case. Rather, the BLUP uses the observed data directly, and the synthetic estimate is used to predict the nonsampled values. With small samples available from most small areas these two estimators will be quite similar. However, when unequal probabilities of selection are used in situations such as populations with highly skewed distributions for Y (e.g., energy consumption per establishment) the two estimators may be quite different if, for example, the sample includes with certainty all of the members of the population with the largest values for Y. It is important to note that the superpopulation (empirical Bayes) form of the synthetic estimator in Equation (6) is asymptotically consistent (as n_{ij} approaches N_{ij}, $\hat{\bar{Y}}_{i..}$ approaches $\bar{Y}_{i..}$), while the traditional synthetic estimator is not.

Gonzalez and Waksberg (1973) derive a general variance estimate for synthetic estimators. They suggest using the average mean squared error (MSE) for all domain estimates and using it show synthetic estimates to be preferable to regional averages for estimating errors in vacancy rates in the 1970 U.S. Census. Their MSE estimate gives only an estimate of average error over all domains; there is no way to derive an estimate of the accuracy of any single small area estimate from this procedure. Despite the shortcomings of this estimator, it makes a large step forward in providing a measure of accuracy for small area estimates that can be applied without assuming the validity of the underlying model. Marker (1995) improves on this estimator by developing a small area-specific MSE that combines a small area-specific variance estimate with an average squared bias term. If the variance is a large component of the MSE, this estimator can identify small areas whose estimates are significantly more accurate than the estimates for other small areas.

To examine the effectiveness of the average MSE as a measure of accuracy, Gonzalez (1973) compares a standard normal distribution to the empirical distribution of the biases (as measured against the 1960 U.S. Census values) of the synthetic estimates in multiples of the average root mean squared error (RMSE). The empirical distribution (when the true total is small) is more concentrated inside one RMSE and beyond two RMSE than is the standard normal distribution. The larger the true total the less obvious this difference from a normal distribution becomes. For the largest estimates the deviations are reversed, with the empirical distribution being flatter than the normal distribution.

Other papers using synthetic estimation for small areas include those by Aliaga and Le (1991), DiGaetano, Waksberg, Mackenzie, and Yaffe (1980), Gonzalez and Hoza (1978), Levy and French (1977), Namekata, Levy, and O'Rourke (1975), Schaible (1980), and Shpiece (1981). DiGaetano et al. and Nemkata et al. both examined using synthetic estimation for the National Health Interview Survey. The Aliaga and Le, DiGaetano et al., Gonzalez and Hoza, and Levy and French papers compared synthetic estimation with other small area techniques. The Levy and French, Namekata et al., Schaible, and Shpiece papers all discussed the impact of the assumptions underlying the synthetic estimator.

Levy (1978) attempts to place synthetic estimation in a multiple regression framework, with the subgroup means being the unknowns estimated from the sample. The independent variables in his regression are the population proportions of subgroups j within small area i. However, use of the population proportions as the independent variables does not produce the synthetic estimator as a regression estimator. As shown below, the independent variables should be indicator variables of the different subgroups. Let each subgroup cross-classification (e.g., 18- to 44-year-old black males) be a separate symptomatic variable, and let $Y_{ijk} = B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_JX_J + e_{ijk}$, where

$$X_j = \begin{cases} 1, & \text{if } Y_{ij'k} \text{ is such that } j' = j \\ 0, & \text{otherwise} \end{cases}$$

and $e_{ijk} \sim N(0, \sigma^2 I)$. The least squares estimate of $\mathbf{B} = [B_1, \ldots, B_J]'$ is then $\hat{\mathbf{B}} = (X'X)^{-1}X'\mathbf{Y}$, where X is an $(n_{..} \times J)$ matrix in which each row is entirely filled with zeros except for a one in the column representing the subgroup to which that element belongs. Sort the sampled units so that all $n_{.1}$ rows of the X matrix with a one in column one are first, all $n_{.2}$ rows with a one in column two are second, etc. Then $(X'X)^{-1} = \text{diag}\{n_{.1}^{-1}, n_{.2}^{-1}, n_{.3}^{-1}, \ldots, n_{.J}^{-1}\}$; $X'\mathbf{Y} = (y_{.1.}, y_{.2.}, y_{.3.}, \ldots y_{.J.})'$; $\hat{\mathbf{B}} = (X'X)^{-1}X'\mathbf{Y} = (\bar{y}_{.1.}, \bar{y}_{.2.}, \bar{y}_{.3.}, \ldots, \bar{y}_{.J.})'$; $\tilde{Y}_{i..} = \Sigma_j \Sigma_k \hat{\mathrm{B}}_j X_j = \Sigma_j N_{ij} \bar{y}_{.j.}$; where diag{} is a diagonal $(J \times J)$ matrix, $\tilde{Y}_{i..}$ is the regression estimate of the total, and N_{ij} denotes the population size of subgroup j in small area i. That is, when the X matrix is composed of indicator variables, it is possible to put synthetic estimation in a multiple regression framework. In particular, synthetic estimation is shown to fit an individual-level regression model of the same form as components of variance, where the small area random effects, v_i, are all set equal to zero.

The estimates produced by the synthetic estimator can be unstable when the subgroup means are themselves unstable. This can occur for two different reasons: the overall sample size may be small, so that the subgroup means are based on small samples; or when the overall sample size is large, the number of variables that impact the estimates, and therefore the number of subgroups, is large, so that the resulting sample size for each subgroup is small. There are two methods for improving the stability of these subgroup means (and hence the synthetic estimates). First, the means can be modeled and predicted from all of the data where the variables that determine the subgroups are used as the independent variables (Elston, Koch, Weissert 1991). The predicted value for each subgroup is then used in the synthetic estimate instead of the sample average (in Equation (5)). To the extent that the model describes the true underlying relationship in the data these predicted values will produce accurate estimates for each small area. The second approach is to place a prior distribution on the subgroup means and combine the sample data with prior information to produce estimates for each subgroup that are optimal conditional on the prior (Marker 1995).

Synthetic estimation uses past census data (sometimes updated from other sources) to determine the subgroup composition of each small area. It does not make any assumptions about historical correlations among variables. It is assumed that the average response for a given subgroup is the same in every small area. The failure of this assumption causes the synthetic estimator to underestimate large deviations from the overall average that occur in some small areas. Three adjustments to the standard synthetic estimator have been proposed (referred to as composite estimators) to make it more sensitive to these deviations and are described below. These adjustments are to combine the synthetic estimator with either sample-based estimates, administrative records, or regression estimates.

2.1.7. Composite estimators

There are several forms of a composite estimator. The first form uses a weighted average of both the design-unbiased sample mean for small area $i, \bar{y}_{1i.}$, and the synthetic estimate, $\bar{y}_{2i.}$, where the weights are inversely proportional to the mean squared errors of the two estimators (Schaible, Brock, Schnack 1977; Schaible 1978a; Schaible 1978b.) The formula for this composite estimator is

$$\hat{\bar{y}}_{3i.} = t_i\bar{y}_{1i.} + (1 - t_i)\hat{\bar{y}}_{2i.} \tag{7}$$

where

$$t_i = \frac{MSE\left(\hat{\bar{y}}_{2i.}\right)}{Var(\bar{y}_{1i.}) + MSE\left(\hat{\bar{y}}_{2i.}\right)}$$

and $\bar{y}_{1i.} = \Sigma_j\Sigma_k w_{ijk}y_{ijk}/\Sigma_j\Sigma_k w_{ijk}$; w_{ijk} is the weight for the kth respondent in subgroup j in small area i; $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, n_{ij}$; $\hat{\bar{y}}_{2i.} = \Sigma_{j=1}^{J}(N_{ij}\bar{y}_{.j.}/N_{i.})$, where N_{ij} is the population of subgroup j in small area i, $\bar{y}_{.j.} = \Sigma_i\Sigma_k w_{ijk}y_{ijk}/\Sigma_i\Sigma_k w_{ijk}$, and $N_{i.}$ is the population of small area i.

To use this composite estimator, the quantity t_i, which is a function of $Var(\bar{y}_{1i.})$ and $MSE(\hat{\bar{y}}_{2i.})$, needs to be estimated. $Var(\bar{y}_{1i.})$ can be estimated directly from the survey data in small area i. However, since in multi-stage samples many small areas contain only one or two sampled primary sampling units (PSUs), these estimated variances can be subject to extremely large variation. Thus, the $Var(\bar{y}_{1i.})$ are sometimes calculated from generalized variance expressions (Wolter, 1985). A method for the calculation of $MSE(\hat{\bar{y}}_{2i.})$ is described in Marker (1995). When no members of the sample come from a particular small area, then the composite estimator is simply equal to the synthetic estimator.

This composite estimator has been used at the National Center for Health Statistics to calculate state level disability estimates (Malec 1993). The method of Fay and Herriot (1979) is a similar composite estimator using sample regression rather than synthetic estimation. Such an estimator has recently been proposed for allocating 6 billion USD annually to counties in the United States to aid school districts with many poor children (Citro et al. 1997). A composite estimator combining sample regression and direct estimators is developed at the county level. A similar composite estimator (but with different covariates) is derived at the state level. The county-level composite is then ratio-adjusted to match the state-level estimator.

Another alternative composite estimator is to combine the synthetic estimate with the regression methods. The percent deviation of the synthetic estimator from its (assumed) true value can be modeled as the dependent variable in a symptomatic regression equation (Levy 1971). This cannot be solved directly since the true values, and therefore the percent deviations, are unknown. Levy suggests estimating the regression coefficients based on collapsed sets of small areas and then using those values to revise the synthetic estimates for each small area. A more direct approach simply incorporates the synthetic estimator, along with the symptomatic variables, as independent variables in the sample regression method (Nicholls 1977; Gonzalez and Hoza 1978). The method of Nicholls and Gonzalez and Hoza is known as synthetic regression. Thus, synthetic regression assumes that any adjustments to the synthetic estimator to account for model failure can be expressed as a linear function of symptomatic variables.

The Schaible composite estimator can be shown to be similar to the components of variance estimator described earlier. It was previously demonstrated that the synthetic estimator can be expressed as a regression on a set of dummy variables. This allows one to rewrite Equation (7) as $\hat{\hat{y}}_{3i.} = t_i\bar{y}_{1i.} + (1 - t_i)BX$ where X is a set of dummy variables for the subgroups used in the synthetic estimator. In the notation of components of variance the sample mean in small area i can be written as a regression estimate plus a small area random effect, v_i. Thus $\hat{\hat{y}}_{3i.} = t_i^*(BX + v_i) + (1 - t_i^*)BX = BX + t_i^* v_i$. If an average MSE across small areas is used for $MSE(\hat{\hat{y}}_{2i.})$ in Equation (7), then this is of the same form as the components of variance estimator.

2.1.8. Structure preserving estimation

Purcell and Kish (1979, 1980) developed a contingency table approach to small area estimation. Their estimates are known as structure preserving estimates (SPREE) and are derived using the techniques of iterative proportional fitting (IPF). IPF for contingency tables is well documented (see Bishop, Fienberg, Holland 1975) and has many important properties such as known convergence to the maximum likelihood estimates for each cell. When applied to small area estimation, the contingency tables typically have three dimensions corresponding to small areas, the categories of the variable being estimated, and the population subgroups, respectively. The initial values for each cell of the contingency table are typically data from a past census. These values are then updated by IPF to correspond to marginal totals that reflect either a recent sample or updated values of symptomatic variables. Purcell and Kish present six versions of SPREE which allow for when the initial data are known for either all cells or only certain subsets.

One limitation on using SPREE occurs if the size of the sample available for updating is small and one of the sampled marginal values is zero. The SPREE estimate of all cells in the row (or column) adding up to that marginal will be zero, regardless of the original census data.

Pham and Thomsen (1988) present a potential application of the SPREE estimator using administrative records which are available for all small areas in Scandinavian registries. The administrative records are used as the symptomatic variables for updating out-of-date employment data.

SPREE estimates can provide improvements over synthetic estimates by reflecting the more complex relationships between subgroups and small areas that are found in the census data used for initial values. However, if the interrelationships have changed since the last census, SPREE will not reflect such changes. Therefore, the utility of SPREE estimates is highly dependent on the availability of recent census data. This methodology is used in Australia where censuses are conducted every five years (see, for example, Feeney 1987; Steel 1988.) SPREE can be written as a log-linear model, but unlike the earlier estimators discussed in this chapter, it cannot be expressed as a linear regression model.

2.2. Empirical Bayes or superpopulation approaches

The next two subsections discuss empirical Bayes and hierarchical Bayes approaches to small area estimation. While the purpose of these sections is not to compare the different philosophical approaches, it is necessary to understand the basic concepts underlying the methods before discussing the actual models.

Many researchers have proposed alternative empirical Bayes approaches to address the lack of sufficient data for direct estimates. These empirical Bayes approaches (also referred to as superpopulation or predictive approaches) assume that the population data are a sample from a larger superpopulation that can be adequately represented by an empirical Bayes model. Optimal estimators are derived dependent on the assumed model. In some of the papers discussed below, an attempt has been made to examine the robustness of such estimators against different types of model failure.

Scott and Smith (1969) present a clear description of the different populations of inference for superpopulation versus traditional survey sampling approaches. They argue that the superpopulation is often what is truly of interest, especially with repeated surveys where the population of inference is recognized to change across the time of the repeated surveys.

A detailed development of the superpopulation approach to survey sampling is given by Cassel, Särndal, and Wretman (1978). They describe various Bayesian approaches, including the concept of exchangeability, but stop short of placing Bayesian prior distributions on their models. The authors of this book have developed a number of empirical Bayes estimators that attempt to balance variance and bias. Comparisons of these estimators are included in Cassel et al. (1987), Särndal and Hidiroglou (1989), and Lundström (1988).

One method for examining the robustness of empirical Bayes small area estimators is to determine how the estimator would change under alternative models. Holt, Smith, and Tomberlin (1979) examine five different sets of assumptions that are commonly used in traditional approaches to small area estimation such as synthetic estimation. For each of these assumptions the optimal model-based estimators of the population total are derived and their MSEs calculated. Next, they address the often-asked but seldom-answered question, ‘‘What if my assumptions are wrong?’’ Their answer is to calculate the biases resulting from using any of these estimators if one of the other four models were instead correct. This procedure provides some indications of how the different estimators behave under less than ideal situations. They suggested moving ‘‘away from the search for a single multipurpose estimator with reasonable average properties that may be poor under certain circumstances toward classes of estimators from which a single estimator is selected for each specific problem.’’

The U.S. Bureau of the Census produces small area estimates using an empirical Bayes technique for estimation of per capita income for 39,000 units of local government (Fay and Herriot 1979). Previously the U.S. Bureau of the Census had imputed county total estimates for unknown values within a county. Fay and Herriot derived two separate estimates for small areas: a sample regression estimate based on census symptomatic data and a direct sample estimate based on the census 20 percent sample in each small area. An average (across small areas) lack of fit for the regression model was calculated, as was the sample variance for the 20 percent sample. Using the inverses of these variance estimates as weights, a combined James-Stein estimator was proposed. In order not to deviate too far from the sample estimate, the final estimate was constrained to deviate no more than one standard error. This estimator was empirically demonstrated to be superior to the previously used county estimates.

2.2.1. Time series models

The empirical Bayes approach can be expanded to borrow strength not only from other small areas but also from previous surveys that include the same variable. Pfeffermann and Burck's (1990) application of this approach involves a survey that is repeated at regular intervals. A time series methodology is used to update the previous survey estimates to the present time period. The Fay-Herriot approach can also be expanded to include random time series effects in addition to fixed small area effects (Rao and Yu, 1994).

Recently there has been increased interest in such time series models for producing small area estimates for labor force surveys. Tiller (1992) modeled U.S. Current Population Survey time series to include trend, seasonal, and irregular components, plus additional explanatory variables. State estimates are then a weighted average of the sample data from the current and all previous months. A simplified version of this model was implemented by the U.S. Bureau of Labour Statistics for 39 states and the District of Columbia. A similar approach has been examined for Canadian unemployment rates (Pfeffermann and Bleuer 1993) where the complex sample rotation patterns were accounted for by separately estimating survey errors for each household panel.

These time series models can also be used to improve trend estimates for small areas. Pfeffermann, Feder, and Signorelli (1998) estimate survey error autocorrelations and separate them out from trend estimates for the Australian labor force survey. They demonstrate that such models can produce better estimates of population values and forecasts of future values.

2.2.2. Models with random and fixed effects

Another expansion of the empirical Bayes framework uses a multiple linear regression model where the deviations from the model are split into random and fixed effects components (Fuller and Harter 1987; Battese, Harter, Fuller 1988.) This allows the expectation to vary from small area to small area. This method can be applied to the situation where one wants to predict Y given X and Z, but Z is measured with error (Fay 1987). A multivariate linear regression model is used to predict Y and Z given X. Fay's example uses state-level estimates of three- and five-person family median income (Z) in predicting four-person family median income (Y). This goes beyond the usual empirical Bayes procedure in that it uses auxiliary data (Z), in addition to the sample, in the estimation process. Datta, Fay, and Ghosh (1991) improve on Fay's procedure by examining empirical Bayes, regression, and two types of hierarchical Bayes estimators for four-person family median income. Ghosh, Nangia, and Kim (1996) expand this further by including time series effects.

The Fuller and Harter model can be extended to allow not only the intercept, but also the regression coefficients, to vary from small area to small area. Holt and Moura (1993a,b) extend the model in this way and also explore the effect of including small area level covariates. Through a series of simulations they examine when the MSE of a small area is likely to be improved by these generalizations and also when the estimators are robust to departures from the assumptions of the models.

For each of these generalizations Holt and Moura produce BLUPs. Introduced by Henderson (1975), BLUPs are the linear models version of empirical Bayes estimators when the model includes random effects. Robinson (1991) examined the relationships between these two types of estimators and showed that when the random effects are normally distributed the BLUPs are equivalent to empirical Bayes estimators. When parametric assumptions are made for the random effects the BLUPs are equivalent to parametric empirical Bayes predictions. Harville (1991) pointed out that while the predicted point estimates produced by BLUPs and empirical Bayes estimators are often the same, they treat variance estimation differently. Both frequently underestimate their mean squared errors, but, since empirical Bayes fits easily into a hierarchical Bayes framework, it is easier to incorporate the extra variation into empirical Bayes estimation. Schaible (1993b) and Harville differentiate between the terms BLUE and BLUP by using BLUE for a model parameter (assumed constant under the model) and BLUP for a finite population total or mean.

Prasad and Rao (1990) develop approximate MSEs, and estimates of the MSEs, for the Fuller and Harter, and Fay and Herriot, estimators. They point out that empirical Bayes estimates of MSE underestimate the true MSE by not accounting for the variation in the estimates of the model-dependent variances. By adjusting for this component of variation, their MSE estimator is unbiased, conditional on the model. Lahiri and Rao (1995) show that this estimator of MSE is still unbiased if Prasad and Rao's assumption of normality is eliminated. Prasad and Rao's methodology can be used to produce model-dependent approximate MSEs for the Fay and Herriot estimator expanded to include random time series effects (Rao and Yu 1994).

Thus a number of authors have proposed empirical Bayes approaches to small area estimation, borrowing strength from other small areas. In recent years this approach has been expanded by borrowing strength across time (Tiller 1992), through multivariate regression (Fay 1987), adding small area fixed effects (Fuller and Harter 1987), and adding small area covariates (Holt and Moura 1993a,b). The next section examines the more general approach of placing a prior distribution on the parameters of the small area model.

2.3. Bayesian approaches

It is important when either examining or using model-dependent approaches to emphasize that a model is not chosen because it is believed to be correct, but rather on the basis of one's belief that it will adequately approximate the truth. When using empirical Bayes models it is, therefore, important to examine robustness to model failure. Alternatively, using a subjective or hierarchical Bayesian approach, it can be assumed that a class of models contains the truth. One assumes a prior distribution that is then updated by the observed sample to derive optimal estimators conditional on the specific set of data that are observed.

From the subjective Bayesian view the finite population $\mathbf{Y} = (Y_1, Y_2, \ldots Y_N)$ is unknown but prior information about $\mathbf{Y}$ is known. This is reflected in a prior subjective probability distribution on the random variables $\mathbf{Y}$ given by $p(\mathbf{Y})$. Inference on $\mathbf{Y}$ proceeds by computing the posterior distribution $p(\mathbf{Y}|(s, y))$, where (s, y) are the sample labels, s, and data, y.

Bernardo, in a note on a paper presented by Dempster (1975), suggests the appropriateness of using a Bayesian (as compared to empirical Bayesian) approach by explaining that any superpopulation model can be expressed as part of the prior specification. He suggests that use of a class of priors would lead to more robust inference. This is a view also expressed by Ericson (1982).

Bayesian improvements to the standard regression and synthetic regression methods appear to be particularly appropriate for a survey that is repeated over time. In such cases a hierarchical model can be developed, at least in part, from the experience gained from previous iterations. Malec (1981) assumes the population (e.g., the persons in a country) can be partitioned into exchangeable (Ericson 1969) sub-areas (e.g., regions). Then the regression estimate can be improved by taking a weighted average of the regional and national estimates. He also assumes that there is exchangeability across time. The composite estimate then includes a sample estimate, past and present regional estimates, and past and present national estimates. Malec and Sedransk (1985) derive Bayes estimators for small areas in a three-stage clustered sample design using categorical auxiliary variables. This methodology has been expanded to incorporate continuous auxiliary variables (Farrell, MacGibbon, Tomberlin 1997), but in an empirical Bayesian framework.

Stroud (1987) discusses a Bayesian procedure when an equal sample size has been selected in all small areas. For unequal sample sizes he suggests two alternatives. First, if there are only a few areas, use numerical integration to derive the Bayesian solution since using empirical Bayes would ‘‘ignore uncertainty (arising from estimates based on only a few areas) concerning the variance of the means.’’ Second, if there is a large number of small areas, he suggests using empirical Bayes techniques since this source of uncertainty (in estimating variances) will be reduced.

In some situations one desires to optimally estimate the mean and variance of a set of parameters (e.g., the distribution of median per capita income across small areas). This is a slightly different problem than the typical attempt to develop a best estimate for particular small areas. Louis (1984) provides a theoretical model and empirical example for this by matching the sample mean and variance to the posterior expected mean and variance. His shrinkage factor is the square root of the Bayes shrinkage and is, therefore, inferior for estimating individual means. Thomsen (1994) and Spjøtvoll and Thomsen (1987), provide an example using this shrinkage factor for Norwegian labor statistics.

The work of Louis can be expanded to include constrained hierarchical Bayes estimators. Ghosh (1992) constrains the posterior expected mean and variance to be able to determine a set of small areas whose values are beyond a certain predetermined cut-off, for example those with median incomes of under 15,000 USD. His constraint on the

Table 1. Summary of small area estimation techniques

	Information used:			
Technique	Total Y_i.. estimated by	Sample Data	Symptomatic variables X_i (examples)	Assumptions
Vital rates	$x_{i(t-1)}\frac{x_{st}}{x_{s(t-1)}}$	No	Births, deaths	Ratio of small area rate to state rate invariant since last census.
Symptomatic regression	$X_i^*\hat{\mathbf{B}}$	No	Rates of change of births, deaths, etc. since last census and between the last two censuses	Regression equation invariant since census before the most recent.
Sample regression	$X_i\hat{\mathbf{B}}$	Yes (Dependent variables at area level)	Rates of change of births, deaths, etc.	Changes in sampled small areas are representative of those in other areas.
Components of variance regression	$X_i\hat{\mathbf{B}} + \hat{v}_i$	Yes (Dependent variable at individual level)	Number of tax forms, crops, etc.	Sampled small areas are representative of other areas.
Synthetic estimation	$\sum_{j=1}^{J} N_{ij}\bar{y}_{.j.}$	Yes (Independent variable)	Indicator variables for each subgroup j	Subgroup means are equal in each small area. Population proportions invariant since most recent census.

Note: x_{st} refers to symptomatic variable in state s (containing small area i) at time t. X_i^* are the symptomatic variables for small area i for a more recent time period than is used to compute the regression coefficients.

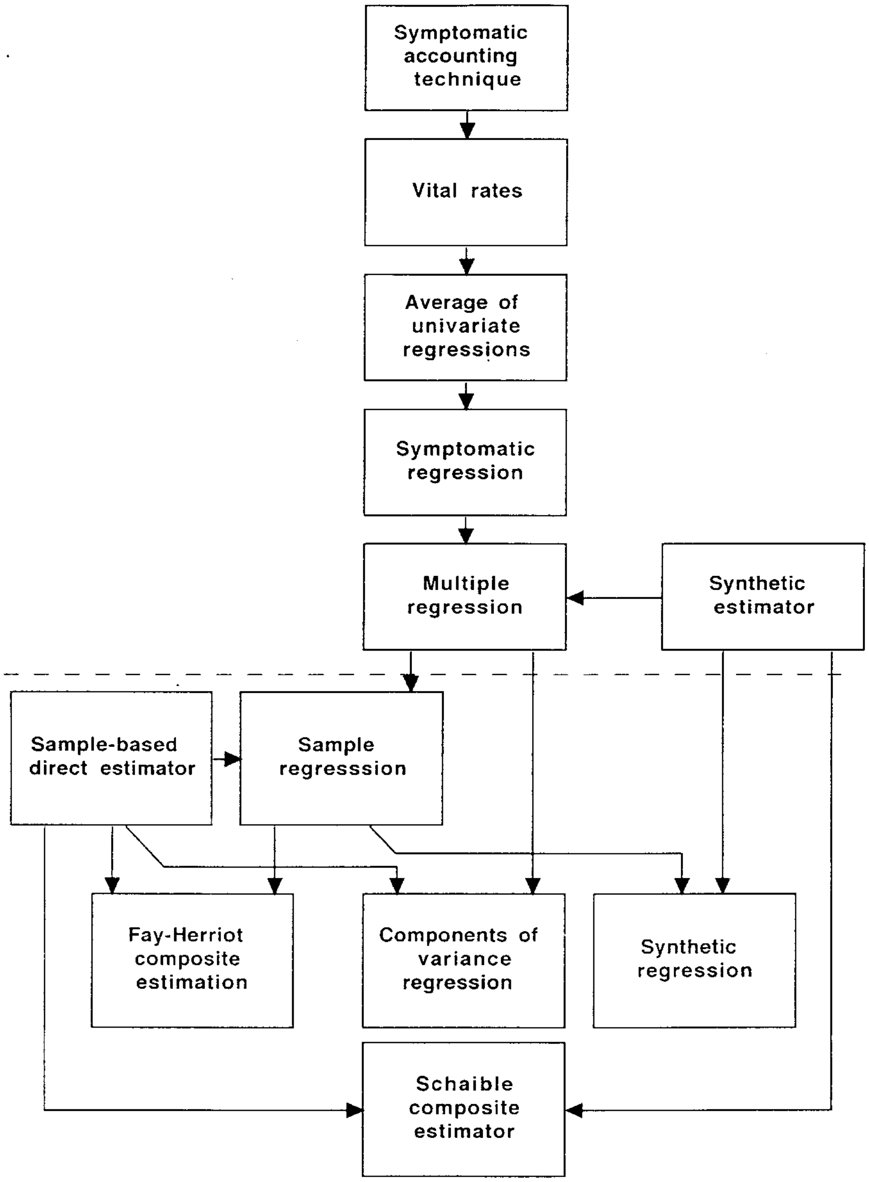
Symptomatic accounting technique
Vital rates
Average of univariate regressions
Symptomatic regression
Multiple regression
Synthetic estimator
Sample-based direct estimator
Sample regresssion
Fay-Herriot composite estimation
Components of variance regression
Synthetic regression
Schaible composite estimator

Fig. 1. Interrelations among small area estimation techniques. Above the dotted line are regression estimators, below the line are combinations of direct and regression estimators

shrinkage is based on the relative magnitudes of the posterior expected variance and the observed variation among the Bayes estimates for the different small areas. Ghosh replaces Louis's assumption of normality with more general assumptions of quadratic loss or posterior linearity (Ericson 1969). This constrained hierarchical Bayes estimator was compared to traditional Bayes and empirical Bayes estimators by Datta, Fay, and Ghosh (1991) and found to provide fairly similar estimates.

Marker (1995) and Marker and Waksberg (1994) use the concept of exchangeability to produce a generalized version of the synthetic estimator. The subgroup means are no longer assumed equal, but rather to be exchangeable with common prior mean and variance. For each subgroup a weighted average is taken of the sample and prior mean. This posterior expectation for each subgroup is then used to produce small area estimates by weighting them according to the subgroup decomposition of the small area. Such an estimator could be of use, for example, when a sample of small areas is selected as part of each cycle of a repeated survey. Historical data would then be available for the mean of subgroups of the population, but not for small area means.

3. Summary of Interrelationships

Section 2 described the interrelationship among the different small area estimation techniques. Many of the estimators can be viewed in a hierarchical structure where each method builds on the others either by relaxing the multiple regression assumptions or including new sources of information. Table 1 summarizes the relationships for vital rates and the four traditional survey sampling regression estimators.

These five regression estimators make very different use of sample data. Vital rates and symptomatic regression do not use sample data, relying on data from the census. Sample regression and components of variance use sample data as part of the dependent variable, the former at the small area level while the latter uses it at the individual level. Synthetic estimation uses sample data at the small area level as the independent variables.

Figure 1 pictorially demonstrates this structure with each arrow above the dotted line pointing toward the more general type of estimator. Bogue (1950) generalized the symptomatic accounting technique into the vital rates technique by using the rates of change since the previous census. Schmitt and Crosetti (1954) provided a multivariate version of the vital rates technique. Vital rates was shown to be a special case of multiple regression on symptomatic variables, and symptomatic regression and the synthetic estimator were also shown to be forms of multiple regression.

A series of small area estimators have been developed using multiple regression, synthetic estimation, and the sample-based direct estimator as building blocks. Ericksen (1973, 1974) suggested generalizing the regression techniques to include sample data at the small area level. This enabled him to loosen the assumptions necessary for symptomatic regression. Fuller and Harter (1987) also included sample data in a (components of variance) regression equation, but at the individual level, and allowed the error term to have two components, one random and the other small area specific. Numerous authors have taken these estimators and placed them in empirical Bayes and hierarchical Bayesian frameworks that allow for direct examination of the appropriate underlying model assumptions.

The sample regression technique can be combined with synthetic estimation to develop the synthetic regression techniques of Levy (1971) or Nicholls (1977) and Gonzalez and Hoza (1978). The sample regression technique may also be combined with the direct sample estimate to derive the composite estimator suggested by Fay and Herriot (1979). Combining the sample estimate with the synthetic estimate results in the composite estimator introduced by Schaible, Brock, and Schnack (1977) and Schaible (1978a, 1978b). Incorporating data from multiple survey cycles allows for the inclusion of time series components in these estimators.

For many small areas there is often no sample available to use in a composite estimator. The sample (or components of variance) regression estimator is always usable (since the model is developed for those areas with sample data and then applied to all areas) but is only effective if a strong linear relationship exists among the available variables.

4. References

Aliaga, A. and Le, T. (1991). Methodology for Small-Area Estimation with DHS Samples. Estadistica, 43, 53–90.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83, 28–36.

Bernardo, J.M. (1975). A Discussion of ''A Subjectivist Look at Robustness'' by Dempster, A.P. and ''Beyond Location Parameters'' by Hempel, F.R. Bulletin of the International Statistical Institute. 46, Book 1, 386–387.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis, Theory and Practice. Cambridge, Massachusetts: MIT Press.

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

Cassel, C.M., Kristiansson, K.E., Råbäck, G., and Wahlström, S. (1987). Using Model-Based Estimation to Improve the Estimate of Unemployment on a Regional Level in the Swedish Labor Force Survey. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 141–159.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: John Wiley and Sons.

Citro, C.F., Cohen, M.L., Kalton, G., and West, K.K. (1997). Small-Area Estimates of School-Age Children in Poverty. Interim Report 1: Evaluation of 1993 County Estimates for Title 1 Allocations. Washington, DC: National Academy Press.

Datta, G.S., Fay, R.E., and Ghosh, M. (1991). Hierarchical and Empirical Multivariate Bayes Analysis in Small Area Estimation. Proceedings of the U.S. Bureau of the Census Annual Research Conference, 63–79.

DiGaetano, R., Waksberg, J., Mackenzie, E., and Yaffe, R. (1980). Synthetic Estimates for Local Areas from the Health Interview Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 46–55.

Elston, J.M., Koch, G.G., and Weissert, W.G. (1990). Regression-Adjusted Small Area Estimates of Functional Dependency in the Noninstitutionalized American Population Age 65 and Over. American Journal of Public Health, 81, 335–43.

Ericksen, E. (1973). Recent Developments in Estimation for Local Areas. Proceedings of the Social Statistics Section, American Statistical Association, 37–41.

Erickson, E. (1974). A Regression Method for Estimating Population Changes of Local Areas. Journal of the American Statistical Association, 69, 867–875.

Ericksen, E. and Kadane, J.B. (1987). Sensitivity Analysis of Local Estimates of Undercount in the 1980 U.S. Census. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 23–45.

Ericson, W. (1969). Subjective Bayesian Models in Sampling Finite Populations. Journal of the Royal Statistical Society B, 31, 195–233.

Ericson, W. (1981). Bayesian Sampling Lecture Notes. Unpublished, University of Michigan.

Ericson, W. (1982). Personal communication.

Ericson, W. (1983). A Bayesian Approach to Regression Estimation in Finite Populations. University of Michigan Technical Report, 120.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997). Empirical Bayes Small-Area Estimation Using Logistic Regression Models and Summary Statistics. Journal of Business and Economic Statistics, 15:1, 101–108.

Fay, R. (1987). Application of Multivariate Regression to Small Domain Estimation. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 91–123.

Fay, R. and Herriot, R. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 74, 269–277.

Feeney, G.A. (1987). The Estimation of the Number of Unemployed at the Small Area Level. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 198–218.

Fuller, W.A. and Harter, R.M. (1987). The Multivariate Components of Variance Model for Small Area Estimation. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 103–123.

Ghangurde, P.D. and Gray, G.B. (1978). Estimation for Small Areas in Household Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 712–715.

Ghosh, M. (1992). Constrained Bayes Estimation with Applications. Journal of the American Statistical Association, 87, 533–540.

Ghosh, M., Nangia, N., and Kim, D.H. (1996). Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach. Journal of the American Statistical Association, 91, 1423–1431.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9:1, 55–93.

González, M. (1973). Use and Evaluation of Synthetic Estimators. Proceedings of the Social Statistics Section, American Statistical Association, 33–36.

González, M. and Hoza, C. (1978). Small Area Estimation with Application to Unemployment and Housing Estimates. Journal of the American Statistical Association, 73, 7–15.

González, M. 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.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. 1. Methods and Applications, Vol. II Theory. New York: John Wiley and Sons.

Harville, D.A. (1991). Discussion of paper by Robinson. Statistical Science, 6, 35–39.

Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction Under a Selection Model. Biometrics, 31, 423–447.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1979). A Model Based Approach to Estimation for Small Subgroups of a Population. Journal of the American Statistical Association, 74, 405–410.

Holt, D. and Moura, F. (1993a). Mixed Models for Making Small Area Estimates. In Kalton, G., Kordos, J., and Platek, R., eds: Small Area Statistics and Survey Designs, 1, Central Statistical Office, Warsaw, Poland, 221–231.

Holt, D. and Moura, F. (1993b). Small Area Estimation Using Multi-Level Models. Proceedings of the Section on Survey Research Methods, American Statistical Association, 21–30.

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

Kish, L. (1987). Discussion. Small Area Statistics. An International Symposium. New York: John Wiley and Sons, 267–271.

Lahiri, P. and Rao, J.N.K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association, 90, 758–766.

Levy, P.S. (1971). The Use of Mortality Data in Evaluating Synthetic Estimates. Proceedings of the Social Statistics Section, American Statistical Association, 328–331.

Levy, P.S. (1978). Small Area Estimation – Synthetic and Other Procedures, 1968–1978. National Institute on Drug Abuse Research Monograph, 24, 4–33.

Levy, P.S. and French, D.K. (1977). Synthetic Estimates of State Health Characteristics Based on the Health Interview Survey. Vital and Health Statistics, 2:75.

Louis, T.A. (1984). Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods. Journal of the American Statistical Association, 79, 393–398.

Lundström, S. (1988). Experiences on Small Domain Estimation Methods at Statistics Sweden. Presented at the NCHS International Symposium on Small Area Statistics, New Orleans, Louisiana.

Malec, D. (1981). Outline of Hierarchical Improvement of Small Area Estimates: U.S. Bureau of the Census. (Mimeographed).

Malec, D. and Sedransk, J. (1985). Bayesian Inference for Finite Population Parameters in Multistage Cluster Sampling. Journal of the American Statistical Association, 80, 879–902.

Malec, D. (1993). Model Based State Estimates from the National Health Interview Survey. Statistical Policy Working Paper 21, Indirect Estimators in Federal Programs. Office of Management and Budget, 8-1–8-25.

Marker, D.A. and Waksberg, J. (1994). Small Area Estimation for the U.S. National Health Interview Survey. Statistics in Transition, 1, 6, 747–768.

Marker, D.A. (1995). Small Area Estimation: A Bayesian Perspective. Unpublished dissertation.

Martin, J.H. and Serow, W.J. (1978). Estimating Demographic Characteristics Using the Ratio-Correlation Method. Demography, 15:2, 223–233.

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, 90:6, 532–538.

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

Nicholls, A. (1977). A Regression Approach to Small Area Estimation. Australian Bureau of Statistics. (Mimeographed).

Otelsberg, J. (1981). Small Area Estimates Based on Concurrent Measures Versus Regression Estimates: The Case of One Family House Construction. Proceedings of the Section on Survey Research Methods, American Statistical Association, 690–709.

Pfeffermann, D. and Barnard, C.H. (1991). Some New Estimators for Small-Area Means with Application to the Assessment of Farmland Values. Journal of Business and Economic Statistics, 9, 73–84.

Pfeffermann, D. and Bleuer, S.R. (1993). Robust Joint Modeling of Labor Force Surveys of Small Areas. Survey Methodology, 19, 149–163.

Pfeffermann, D. and Burck, L. (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. Survey Methodology, 16, 217–238.

Pfeffermann, D., Feder, M., and Signorelli, D. (1999). Estimation of Autocorrelations of Survey Errors With Application to Trend Estimation in Small Areas. Journal of Business and Economic Statistics, forthcoming.

Pham, D.Q. and Thomsen, I. (1988). Small Area Estimation Possibilities and Limitations. Presented at the NCHS International Symposium on Small Area Statistics, New Orleans, Louisiana.

Prasad, N.E. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association, 85, 163–170.

Purcell, N. and Kish, L. (1979). Estimation for Small Domains. Biometrics, 35, 365–384.

Purcell, N. and Kish, L. (1980). Postcensal Estimates for Local Areas. International Statistical Review, 48, 3–18.

Pursell, D.E. (1970). Improving Population Estimates with the use of Dummy Variables. Demography, 7:1, 87–91.

Rao, J.N.K. and Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data. Canadian Journal of Statistics, 22:4, 511–528.

Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. Annals of Mathematical Statistics, 35, 1–20.

Robinson, G.K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science, 6, 15–32.

Särndal, C.E. and Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Analysis. Journal of the American Statistical Association, 84, 266–275.

Schaible, W.L. (1978a). A Composite Estimator for Small Area Statistics. National Institute on Drug Abuse Research Monograph, 24, 36–62.

Schaible, W.L. (1978b). Choosing Weights for Composite Estimators for Small Area Statistics. Proceedings of the Section on Survey Research Methods, American Statistical Association, 741–746.

Schaible, W.L. (1980). A Discussion of ‘‘Synthetic Estimates for Local Areas from the Health Interview Survey’’ by DiGaetano, R., Waksberg, J., Mackenzie, E., and Yaffe, R. Proceedings of the Section on Survey Research Methods, American Statistical Association, 56.

Schaible, W.L. (1993a). Use of Small Area Estimators in U.S. Federal Programs. In

Kalton, G., Kordos, J., and Platek, R., eds: Small Area Statistics and Survey Designs, 1. Central Statistical Office, Warsaw, Poland, 221–231.

Schaible, W.L. (1993b). Personal communication.

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. 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 for Estimating Post-Censal Population. Land Economics, 30:3, 279–281.

Scott, A. and Smith, T.M.F. (1969). Estimation in Multi-Stage Surveys. Journal of the American Statistical Association, 64, 830–840.

Shpiece, M.R. (1981). The Use of Synthetic Estimation in Estimating the Elderly Population in Need: A Study and Comments. Proceedings of the Social Statistics Section, American Statistical Association, 267–272.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and Strategies for Small Area Data. Survey Methodology, 20:1, 3–22.

Spjøtvoll, E. and Thomsen, I. (1987). Application of Some Empirical Bayes Methods to Small Area Statistics. Bulletin of the International Statistical Institute, 2, 435–449.

Steel, D. (1988). Approaches to Small Area Estimation at the Australian Bureau of Statistics. Presented at the NCHS International Symposium on small Area Statistics, New Orleans, Louisiana.

Stroud, T.W.F. (1987). Bayes and Empirical Bayes Approaches to Small Area Estimation. Small Area Statistics, An International Symposium. New York: John Wiley and Sons, 124–137.

Thomsen, I. (1994). A discussion of ‘‘Small Area Estimation: An Appraisal’’ by Ghosh, M. and Rao, J.N.K. Statistical Science, 9:1, 89–90.

Tiller, R.B. (1992). Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey. Journal of Official Statistics, 8, 149–166.

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

U.S. Department of Commerce, Bureau of the Census (1980). Population and Per Capita Money Income Estimates for Local Areas: Detailed Methodology and Evaluation. Current Population Reports, Population Estimates and Projections, Series P-25, 699.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

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 1996
Revised August 1998