소지역통계 (Small Area Statistics)
자료 모음집

1999. 12

통 계 청

목 차

[Statistical Science]

[Journal of the American Statistical Association]

[Survey Methodology]

[The Annals of Statistics]

[52nd ISI Invited Papers]

Statistical Science
1994, Vol. 9, No. 1, 55–93

Small Area Estimation: An Appraisal

M. Ghosh and J. N. K. Rao

Abstract. Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. It is now widely recognized that direct survey estimates for small areas are likely to yield unacceptably large standard errors due to the smallness of sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. This has led to the development of alternative methods such as synthetic, sample size dependent, empirical best linear unbiased prediction, empirical Bayes and hierarchical Bayes estimation. The present article is largely an appraisal of some of these methods. The performance of these methods is also evaluated using some synthetic data resembling a business population. Empirical best linear unbiased prediction as well as empirical and hierarchical Bayes, for most purposes, seem to have a distinct advantage over other methods.

Key words and phrases: Borrowing strength, demographic methods, empirical Bayes, empirical best linear unbiased prediction, hierarchical Bayes, synthetic estimation

1. INTRODUCTION

The terms "small area" and "local area" are commonly used to denote a small geographical area, such as a county, a municipality or a census division. They may also describe a "small domain," i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area. In this paper, we use these terms interchangeably.

The use of small area statistics originated several centuries ago. Brackstone (1987) mentions the existence of such statistics in 11th century England and 17th century Canada. Many other countries may well have similar early histories. However, these early small area statistics were all based either on a census or on administrative records aiming at complete enumeration.

For the past few decades, sample surveys, for most purposes, have taken the place of complete enumeration or census as a more cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. Sample survey data certainly can be used to derive reliable estimators of totals and means for large areas or domains. However, the usual direct survey estimators for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide specific accuracy at a much higher level of aggregation than that of small areas. Thus, until recently, the use of survey data in developing reliable small area statistics, possibly in conjunction with the census and administrative data, has received very little attention.

Things have changed significantly during the last few years, largely due to a growing demand for reliable small area statistics from both the public and private sectors. These days, in many countries including the United States and Canada, there is "increasing government concern with issues of distribution, equity and disparity" (Brackstone, 1987). For example, there may exist geographical subgroups within a given population that are far below the average in certain respects, and need definite upgrading. Before taking remedial action, there is a need to identify such regions, and accordingly, one must have statistical data at the relevant geographical levels. Small area statistics are also needed

M. Ghosh is Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611-2049. J. N. K. Rao is Professor of Statistics, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

in the apportionment of government funds, and in regional and city planning. In addition, there are demands from the private sector since the policy-making of many businesses and industries relies on local socio-economic conditions. Thus, the need for small area statistics can arise from diverse sources.

Demands of the type described above could not have been met without significant advances in statistical data processing. Fortunately, with the advent of high-speed computers, fast processing of large data sets made feasible the provision of timely data for small areas. In addition, several powerful statistical methods with sound theoretical foundation have emerged for the analysis of local area data. Such methods "borrow strength" from related or similar small areas through explicit or implicit models that connect the small areas via supplementary data (e.g., census and administrative records). However, these methods are not readily available in a package to the user, and a unified presentation which compares and contrasts the competing methods has not been attempted before.

Earlier reviews on the topic of small area estimation focussed on demographic methods for population estimation in post-censual years. Morrison (1971) covers the pre-1970 period very well, including a bibliography. National Research Council (1980) provides detailed information as well as a critical evaluation of the Census Bureau's procedures for making post-censual estimates of the population and per capita income for local areas. Their document was the report of a panel on small-area estimates of population and income set up by the Committee on National Statistics at the request of the Census Bureau and the Office of Revenue Sharing of the U.S. Department of Treasury. This document also assessed the "levels of accuracy of current estimates in light of the uses made of them and of the effect of potential errors on these uses." Purcell and Kish (1979) review demographic methods as well as statistical methods of estimation for small domains. An excellent review provided by Zidek (1982) introduces a criterion that can be used to evaluate the relative performance of different methods for estimating the populations of local areas. McCullagh and Zidek (1987) elaborate this criterion more fully. Statistics Canada (1987) provides an overview and evaluation of the population estimation methods used in Canada.

Prompted by the growing demand for reliable small area statistics, several symposia and workshops were also organized in recent years, and some of the proceedings have also been published: National Institute on Drug Abuse, Princeton Conference (see National Institute on Drug Abuse, 1979), International Symposium on Small Area Statistics, Ottawa [see Platek et al. (1987) for the invited papers and Platek and Singh (1986) for the contributed papers presented at the symposium]; International Symposium on Small Area Statistics, New Orleans, 1988, organized by the National Center for Health Statistics; Workshop on Small Area Estimates for Military Personnel Planning, Washington, D.C., 1989, organized by the Committee on National Statistics; International Scientific Conference on Small Area Statistics and Survey Designs, Warsaw, Poland, 1992, (see Kalton, Kordos and Platek, 1993). The published proceedings listed above provide an excellent collection of both theoretical and application papers.

Reviews by Rao (1986) and Chaudhuri (1992) cover more recent techniques as well as traditional methods of small area estimation. Schaible (1992) provides an excellent account of small area estimators used in U.S. Federal programs (see NTIS, 1993, for a full report prepared by the Subcommittee on Small Area Estimation of the Federal Committee on Statistical Methodology, Office of Management and Budget).

The present article considerably updates earlier reviews by introducing several recent techniques and evaluating them in the light of practical considerations. Particularly noteworthy among the newer methods are the empirical Bayes (EB), hierarchical Bayes (HB) and empirical best linear unbiased prediction (EBLUP) procedures which have made significant impact on small area estimation during the past decade. Before discussing these methods in the sequel, it might be useful to mention a few important applications of small area estimation methods as motivating examples.

As our first example, we cite the Federal-State Cooperative Program (FSCP) initiated by the U.S. Bureau of the Census in 1967 (see National Research Council, 1980). A basic goal of this program was to provide high-quality, consistent series of county population estimates with comparability from area to area. Forty-nine states (with the exception of Massachusetts) currently participate in this program, and their designated agencies work together with the Census Bureau under this program. In addition to county estimates, several members of the FSCP now produce subcounty estimates as well. The FSCP plays a key role in the Census Bureau's post censual estimation program as the FSCP contacts provide the bureau a variety of data that can be used in making post censual population estimates. Considerable methodological research on small area population estimation is being conducted in the Census Bureau.

Our second example is taken from Fay and Herriot (1979) whose objective was to estimate the per

capita income (PCI) for several small places. The U.S. Census Bureau was required to provide the Treasury Department with the PCI estimates and other statistics for state and local governments receiving funds under the General Revenue Sharing Program. These statistics were then used by the Treasury Department to determine allocations to the local governments within the different states by dividing the corresponding state allocations. Initially, the Census Bureau determined the current estimates of PCI by multiplying the 1970 census estimates of PCI in 1969 (based on a 20 percent sample) by ratios of an administrative estimate of PCI in the current year and a similarly derived estimate for 1969. The bureau then confronted the problem that among the approximately 39,000 local government units about 15,000 were for places having fewer than 500 persons in 1970. The sampling errors in the PCI estimates for such small places were large: for a place of 500 persons the coefficient of variation was about 13 percent while it increased to about 30 percent for a place of 100 persons. Consequently, the Bureau initially decided to set aside the census estimates for these small areas and use the corresponding county PCI estimates in their place. This solution proved unsatisfactory, however, in that the census estimates of PCI for a large number of small places differed significantly from the corresponding county estimates, after taking account of the sampling errors. Fay and Herriot (1979) suggest better estimates based on the EB method and present empirical evidence that these have average error smaller than either the census sample estimates or the county averages. The proposed estimate for a small place is a weighted average of the census sample estimate and a "synthetic" estimate obtained by fitting a linear regression equation to the sample estimates of PCI using as independent variables the corresponding county averages, tax-return data for 1969 and data on housing from the 1970 census. The Fay-Herriot method was adopted by the Census Bureau in 1974 to form updated estimates of PCI for small places. Section 4 discusses the Fay-Herriot model and similar models for other purposes, all involving linear regression models with random small area effects.

Our third example refers to the highly debated and controversial issue of adjusting for population undercount in the 1980 U.S. Census. Every tenth year since 1790 a census has been taken to count the U.S. population. The census provides the population count for the whole country as well as for each of the 50 states, 3000 counties and 39,000 civil divisions. These counts are used by the Congress for apportioning funds, amounting to about 100 billion dollars a year during the early 1980s, to the different state and local governments.

It is now widely recognized that complete coverage is impossible. In 1980, vast sums of money and intellectual resources were expended by the U.S. Census Bureau on the reduction of non-coverage. Despite this, there were complaints of undercounts by several major cities and states for their respective areas, and indeed New York State filed a lawsuit against the Census Bureau in 1980 demanding the Bureau to revise its count for that state.

An undercount is the difference between omissions and erroneous inclusions in the census, and it is typically positive. In New York State's law suit against the Census Bureau, E.P. Ericksen and J.B. Kadane, among other statisticians, appeared as the plaintiff's expert witnesses. They proposed using weighted averages of sample estimates and synthetic regression estimates of the 1980 Census undercount, similar to those of Fay and Herriot (1979) for PCI, to arrive at the adjusted population counts of the 50 states and the 16 large cities, including the State of New York and New York City. The sample estimates are obtained from a Post Enumeration Survey. Their general philosophy on the role of adjustment as well as the explicit regression models used for obtaining the regression estimates are documented in Ericksen and Kadane (1985) and Ericksen, Kadane and Tukey (1989). These authors also suggest using the regression equation for areas where no sample data are available. As a historical aside, we may point out here that the regression method for improving local area estimates was first used by Hansen, Hurwitz and Madow (1953, pages 483–486), but its recent popularity owes much to Ericksen (1974).

While the Ericksen-Kadane proposal was applauded by many as the first serious attempt towards adjustment of Census undercount, it has also been vigorously criticized by others (see, e.g., the discussion of Ericksen and Kadane, 1985). In particular, Freedman and Navidi (1986, 1992) criticized them for not validating their model and for not making their assumptions explicit. They also raise several other technical issues, including the effect of large biases and large sampling errors in the sample estimates. Ericksen and Kadane (1987, 1992), Cressie (1989, 1992), Isaki et al. (1987) and others address these difficulties, but clearly further research is needed. Researchers within and outside the U.S. Census Bureau are currently studying various models for census undercount and the properties of the resulting estimators and associated measures of uncertainty using the EBLUP, EB, HB and related approaches.

Our fourth example, taken from Battese, Harter

and Fuller (1988), concerns the estimation of areas under corn and soybeans for each of 12 counties in North-Central Iowa using farm-interview data in conjunction with LANDSAT satellite data. Each county was divided into area segments, and the areas under corn and soybeans were ascertained for a sample of segments by interviewing farm operators; the number of sample segments in a county ranged from 1 to 6. Auxiliary data in the form of numbers of pixels (a term used for "picture elements" of about 0.45 hectares) classified as corn and soybeans were also obtained for all the area segments, including the sample segments, in each county using the LANDSAT satellite readings. Battese, Harter and Fuller (1988) employ a "nested error regression" model involving random small area effects and the segment-level data and then obtain the EBLUP estimates of county areas under corn and soybeans using the classical components of variance approach (see Section 5). They also obtain estimates of mean squared error (MSE) of their estimates by taking into account the uncertainty involved in estimating the variance components. Datta and Ghosh (1991) apply the HB approach to these data and show that the two approaches give similar results.

Our final example concerns the estimation of mean wages and salaries of units in a given industry for each census division in a province using gross business income as the only auxiliary variable with known population means (see Särndal and Hidiroglou, 1989). This example will be used in Section 6 to compare and evaluate, under simple random sampling, several competing small area estimators discussed in this paper, treating the census divisions as small areas. We were able to compare the actual errors of the different small area estimators since the true mean wages and salaries for each small area are known.

The outline of the paper is as follows. Section 2 gives a brief account of classical demographic methods for local estimation of population and other characteristics of interest in post-censual years. These methods use current data from administrative registers in conjunction with related data from the latest census. Section 3 provides a discussion of traditional synthetic estimation and related methods under the design-based framework. Two types of small area models that include random area-specific effects are introduced in Section 4. In the first type, only area specific auxiliary data, related to parameters of interest, are available. In the second type of models, element-specific auxiliary data are available for the population elements; and the variable of interest is assumed to be related to these variables through a nested error regression model. We present the EBLUP, EB and HB approaches to small area estimation in Section 5 in the context of basic models given in Section 4. Both point estimation and measurement of uncertainty associated with the estimators are studied. Section 6 compares the performances of several competing small area estimators using sample data drawn from a synthetic population resembling the business population studied by Särndal and Hidiroglou (1989). In Section 7, we focus on special problems that may be encountered in implementing model-based methods for small area estimation. In particular, we give a brief account of model diagnostics for the basic models of Section 4 and of constrained estimation. Various extensions of the basic models are also mentioned in this section. Finally, some concluding remarks are made in Section 8.

The scope of our paper is limited to methods of estimation for small areas; but the development and provision of small area statistics involves many other issues, including those related to sample design and data development, organization and dissemination. Brackstone (1987) gives an excellent account of these issues in the context of Statistics Canada's Small Area Data Program. Singh, Gambino and Mantel (1992) highlight the need for developing an overall strategy that includes planning, designing and estimation stages in the survey process.

2. DEMOGRAPHIC METHODS

As pointed out earlier, demographers have long been using a variety of methods for local estimation of population and other characteristics of interest in post-censual years. Purcell and Kish (1980) categorize these methods under the general heading of Symptomatic Accounting Techniques (SAT). Such techniques utilize current data from administrative registers in conjunction with related data from the latest census. The diverse registration data used in the U.S. include "symptomatic" variables, such as the numbers of births and deaths, of existing and new housing units and of school enrollments whose variations are strongly related to changes in population totals or in its components. The SAT methods studied in the literature include the Vital Rates (VR) method (Bogue, 1950), the composite method (Bogue and Duncan, 1959), the Census Component Method II (CM-II) (U.S. Bureau of the Census, 1966), and the Administrative Records (AR) method (Starsinic, 1974), and the Housing Unit (HU) method (Smith and Lewis, 1980).

The VR method uses only birth and death data, and these are used as symptomatic variables rather than as components of population change. First, in a given year, say t, the annual number of births,

b_t, and deaths, d_t, are determined for a local area. Next the crude birth and death rates, r_{1t} and r_{2t}, for that local area are estimated by

$$r_{1t} = r_{10}(R_{1t}/R_{10}), \qquad r_{2t} = r_{20}(R_{2t}/R_{20}),$$

where r_{10} and r_{20} respectively denote the crude birth and death rates for the local area in the latest census year ($t = 0$) while $R_{1t}(R_{2t})$ and $R_{10}(R_{20})$ respectively denote the crude birth (death) rates in the current and census years for a larger area containing the local area. The population P_t for the local area at year t is then estimated by

$$P_t = \tfrac{1}{2}(b_t/r_{1t} + d_t/r_{2t}).$$

As pointed out by Marker (1983), the success of the VR method depends heavily on the validity of the assumption that the ratios r_{1t}/r_{10} and r_{2t}/r_{20} for the local area are approximately equal to the corresponding ratios, R_{1t}/R_{10} and R_{2t}/R_{20}, for the larger area. Such an assumption is often questionable, however.

The composite method is an extension of the VR method that sums independently computed age-sex-race specific estimates based on births, deaths and school enrollments (see Zidek, 1982, for details).

The CM-II method takes account of net migration unlike the previous methods. Denoting the net migration in the local area during the period since the last census as m_t, an estimate of P_t is given by

$$P_t = P_0 + b_t - d_t + m_t,$$

where P_0 is the population of the local area in the census year $t = 0$. In the U.S., the net migration is further subdivided into military and civilian migration. The former is readily obtainable from administrative records while the CM-II estimates civilian migration from school enrollments. The AR method, on the other hand, estimates the net migration from records for individuals as opposed to collect units like schools (see Zidek, 1982, for details).

The HU method expresses P_t as

$$P_t = (H_t)(PPH_t) + GQ_t,$$

where H_t is the number of occupied housing units at time t, PPH_t is the average number of persons per housing unit at time t and GQ_t is the number of persons in group quarters at time t. The quantities H_t, PPH_t and GQ_t all need to be estimated. Smith and Lewis (1980) report different methods of estimating these quantities.

As pointed out by Marker (1983), most of the estimation methods mentioned above can be identified as special cases of multiple linear regression.

Regression-symptomatic procedures also use multiple linear regression for estimating local area populations utilizing symptomatic variables as independent variables in the regression equation. Two such procedures are the ratio-correlation method and the difference-correlation method. Briefly, the former method is as follows: Let 0, 1 and $t(> 1)$ denote two consecutive census years and the current year, respectively. Also, let $P_{i\alpha}$ and $S_{ij\alpha}$ be the population and the value of the jth symptomatic variable for the ith local area ($i = 1, \ldots, m$) in the year $\alpha(= 0, 1, t)$. Further, let $p_{i\alpha} = P_{i\alpha}/\Sigma_i P_{i\alpha}$ and $s_{ij\alpha} = S_{ij\alpha}/\Sigma_i S_{ij\alpha}$ be the corresponding proportions, and write $R_i' = p_{i1}/p_{i0}$, $R_i = p_{it}/p_{i1}$, $r_{ij}' = s_{ij1}/s_{ij0}$ and $r_{ij} = s_{ijt}/s_{ij1}$. Using the data $(R_i', r_{i1}', \ldots, r_{ip}'; i = 1, \ldots, m)$ and multiple regression, we first fit

$$R_i' = \hat{\beta}_0' + \hat{\beta}_1' r_{i1}' + \ldots + \hat{\beta}_p' r_{ip}', \tag{2.1}$$

where $\hat{\beta}$s are the estimated regression coefficients that link the change, R_i', in the population proportions between the two census years to the corresponding changes, r_{ij}', in the proportions for the symptotmatic variables. Next the changes, R_i, in the post censual period are predicted as

$$\tilde{R}_i = \hat{\beta}_0' + \hat{\beta}_1' r_{i1} + \ldots + \hat{\beta}_p' r_{ip},$$

using the known changes, r_{ij}, in the symptomatic proportions in the post censual period and the estimated regression coefficients. Finally, the current population counts, P_{it}, are estimated as

$$\tilde{P}_{it} = \tilde{R}_i p_{i1} \left(\sum_i P_{it} \right),$$

where the total current count, $\Sigma_i P_{it}$, is ascertained from other sources. In the difference-correlation method, differences between the proportions at the two pairs of time points, $(0, 1)$ and $(1, t)$, are used rather than their ratios.

The regression-symptomatic procedures described above use the regression coefficients, $\hat{\beta}_j'$, in the last intercensual period, but significant changes in the statistical relationship can lead to errors in the current postcensal estimates. The sample-regression method (Ericksen, 1974) avoids this problem by using sample estimates of R_i to establish the current regression equation. Suppose sample estimates of R_i are available for k out of m local areas, say $\hat{R}_1, \ldots, \hat{R}_k$. Then one fits the regression equation

$$\hat{R}_i = \hat{\beta}_0 + \hat{\beta}_1 r_{i1} + \ldots + \hat{\beta}_p r_{ip}$$

to the data $(\hat{R}_i, r_{i1}, \ldots, r_{ip})$ from the k sampled areas, instead of (2.1); and then obtains the sample-regression estimators, $\hat{R}_{i(\text{reg})}$, for all the areas using the known symptomatic ratios r_{ij} $(i = 1, \ldots, m)$:

$$\hat{R}_{i(\text{reg})} = \hat{\beta}_0 + \hat{\beta}_1 r_{i1}, + \ldots + \hat{\beta}_p r_{ip}.$$

Using 1970 census data and sample data from the Current Population Survey (CPS), Ericksen (1974) has shown that the reduction of mean error is slight compared to the ratio-correlation method but that of large errors (10% or greater) is more substantial. The success of Ericksen's method depends largely on the size and quality of the samples, the dynamics of the regression relationships and the nature of the variables.

3. SYNTHETIC AND RELATED ESTIMATORS

Gonzalez (1973) describes synthetic estimates as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates." The National Center for Health Statistics (1968) first used synthetic estimation to calculate state estimates of long and short term physical disabilities from the National Health Interview Survey data. This method is traditionally used for small area estimation, mainly because of its simplicity, applicability to general sampling designs and potential of increased accuracy in estimation by borrowing information from similar small areas. We now give a brief account of synthetic estimation and related methods, under the design-based framework.

3.1 Synthetic Estimation

Suppose the population is partitioned into large domains g for which reliable direct estimators, $\hat{Y}'_{\cdot g}$, of the totals, $Y_{\cdot g}$, can be calculated from the survey data; the small areas, i, may cut across g so that $Y_{\cdot g} = \Sigma_i Y_{ig}$, where Y_{ig} is the total for cell (i,g). We assume that auxiliary information in the form of totals, X_{ig}, is also available. A synthetic estimator of small area total $Y_i = \Sigma_g Y_{ig}$ is then given by

$$\hat{Y}_i^S = \sum_g (X_{ig}/X_{\cdot g})\hat{Y}'_{\cdot g}, \tag{3.1}$$

where $X_{\cdot g} = \Sigma_i X_{ig}$ (Purcell and Linacre, 1976; Ghangurde and Singh, 1977). The estimator (3.1) has the desirable consistency property that $\Sigma_i \hat{Y}_i^S$ equals the reliable direct estimator $\hat{Y}' = \Sigma_g \hat{Y}'_{\cdot g}$ of the population total Y, unlike the original estimator proposed by the National Center for Health Statistics (1968) which uses the ratio $X_{ig}/\Sigma_g X_{ig}$ instead of $X_{ig}/X_{\cdot g}$.

The direct estimator $\hat{Y}'_{\cdot g}$ used in (3.1) is typically a ratio estimator of the form

$$\hat{Y}'_{\cdot g} = \left[\left(\sum_{\ell \in s_{\cdot g}} w_\ell y_\ell\right) \Big/ \left(\sum_{\ell \in s_{\cdot g}} w_\ell x_\ell\right)\right] X_{\cdot g} = (\hat{Y}_{\cdot g}/\hat{X}_{\cdot g})X_{\cdot g},$$

where $s_{\cdot g}$ denotes the sample in the large domain g and w_ℓ is the sampling weight attached to the ℓth element. For this choice, the synthetic estimator (3.1) reduces to $\hat{Y}_i^S = \Sigma_i X_{ig}(\hat{Y}_{\cdot g}/\hat{X}_{\cdot g})$.

If $\hat{Y}'_{\cdot g}$ is approximately design-unbiased, the design-bias of $\hat{Y}_i^S$ is given by

$$E(\hat{Y}_i^S) - Y_i \doteq \sum_g X_{ig}(Y_{\cdot g}/X_{\cdot g} - Y_{ig}/X_{ig}),$$

which is not zero unless $Y_{ig}/X_{ig} = Y_{\cdot g}/X_{\cdot g}$ for all g. In the special case where the auxiliary information X_{ig} equals the population count N_{ig}, the latter condition is equivalent to assuming that the small area means $\bar{Y}_{ig}$ in each group g equal the overall group mean, $\bar{Y}_{\cdot g}$. Such an assumption is quite strong, and in fact synthetic estimators for some of the areas can be heavily biased in the design-based framework.

It follows from (3.1) that the design-variance of $\hat{Y}_i^S$ will be small since it depends only on the variances and covariances of the reliable estimators $\hat{Y}'_{\cdot g}$. The variance of $\hat{Y}_i^S$ is readily estimated, but it is more difficult to estimate the MSE of $\hat{Y}_i^S$. Under the assumption $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$, where $\hat{Y}_i$ is a direct, unbiased estimator of Y_i, an approximately unbiased estimator of MSE is given by

$$\text{mse}(\hat{Y}_i^S) = (\hat{Y}_i^S - \hat{Y}_i)^2 - v(\hat{Y}_i). \tag{3.2}$$

Here $v(\hat{Y}_i)$ is a design-unbiased estimator of variance of $\hat{Y}_i$. The estimators (3.2), however, are very unstable. Consequently, it is customary to average these estimators over i to get a stable estimator of MSE (Gonzalez, 1973), but such a global measure of uncertainty can be misleading. Note that the assumption $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$ may be realistic in practice since $\hat{Y}_i^S$ is much less variable than $\hat{Y}_i$.

Nichol (1977) proposes to add the synthetic estimate, $\hat{Y}_i^S$, as an additional independent variable in the sample-regression method. This method, called the combined synthetic-regression method, showed improvement, in empirical studies, over both the synthetic and sample-regression estimates.

Chambers and Feeney (1977) and Purcell and Kish (1980) propose structure preserving estimation (SPREE) as a generalization of synthetic estimation in the sense it makes a fuller use of reliable direct estimates. SPREE uses the well-known method of iterative proportional fitting of margins in a multiway table, where the margins are direct estimates.

3.2 Composite Estimation

A natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators. Such composite estimators may be written as

$$\text{(3.3)} \qquad \hat{Y}_i^C = w_i \hat{Y}_{1i} + (1 - w_i)\hat{Y}_{2i},$$

where $\hat{Y}_{1i}$ is a direct estimator, $\hat{Y}_{2i}$ is an indirect estimator and w_i is a suitably chosen weight ($0 \le w_i \le 1$). For example, the unbiased estimator $\hat{Y}_i$ may be chosen as $\hat{Y}_{1i}$, and the synthetic estimator $\hat{Y}_i^S$ as $\hat{Y}_{2i}$. Many of the estimators proposed in the literature, both design-based and model-based, have the form (3.3). Section 5 gives such estimators under realistic small area models that account for area-specific effects. In this subsection, we mainly focus on the determination of weights, w_i, in the design-based framework using $\hat{Y}_{1i} = \hat{Y}_i$ and $\hat{Y}_{2i} = \hat{Y}_i^S$.

Optimal weights, $w_i(\text{opt})$, may be obtained by minimising the MSE of $\hat{Y}_i^C$ with respect to w_i assuming $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$:

$$\text{(3.4)} \quad w_i(\text{opt}) = \text{MSE}\,(\hat{Y}_i^S)/[\text{MSE}\,(\hat{Y}_i^S) + V(\hat{Y}_i)].$$

The optimal weight (3.4) may be estimated by substituting the estimator mse $(\hat{Y}_i^S)$ given in (3.2) for the numerator and $(\hat{Y}_i^S - \hat{Y}_i)^2$ for the denominator, but the resulting weights can be very unstable. Schaible (1978) proposes an "average" weighting scheme based on several variables to overcome this difficulty, noting that the composite estimator is quite robust to deviations from $w_i(\text{opt})$. Another approach (Purcell and Kish, 1979) uses a common weight, w, and then minimizes the average MSE, i.e., $m^{-1}\Sigma_i$ MSE $(\hat{Y}_i^C)$, with respect to w. This leads to estimated weight of the form

$$\text{(3.5)} \qquad \hat{w}(\text{opt}) = 1 - \sum_i v(\hat{Y}_i) \Big/ \sum_i (\hat{Y}_i^S - \hat{Y}_i)^2.$$

If the variances of $\hat{Y}_i$'s are approximately equal, then we can replace $v(\hat{Y}_i)$ by the average $\bar{v} = \Sigma_i v(\hat{Y}_i)/m$ in which case (3.5) reduces to James-Stein type weight:

$$\hat{w}(\text{opt}) = 1 - m\bar{v} \Big/ \sum_i (\hat{Y}_i^S - \hat{Y}_i)^2.$$

The choice of a common weight, however, is not reasonable if the individual variances, $V(\hat{Y}_i)$, vary considerably. Also, the James-Stein estimator can be less efficient than the direct estimator, $\hat{Y}_i$, for some individual areas if the small areas that are pooled are not "similar" (C.R. Rao and Shinozaki, 1978).

Simple weights, w_i, that depend only on the domain counts or the domain totals of a covariate x have also been proposed in the literature. For example, Drew, Singh and Choudhry (1982) propose the sample size dependent estimator which uses the weight

$$\text{(3.6)} \qquad w_i(D) = \begin{cases} 1, & \text{if } \hat{N}_i \ge \delta N_i, \\ \hat{N}_i/(\delta N_i), & \text{otherwise,} \end{cases}$$

where $\hat{N}_i$ is the direct, unbiased estimator of the known domain population size N_i and δ is subjectively chosen to control the contribution of the synthetic estimator. This estimator with $\delta = 2/3$ and a generalized regression synthetic estimator replacing the ratio synthetic estimator $\hat{Y}_i^S$ is currently being used in the Canadian Labour Force Survey to produce domain estimates. Särndal and Hidiroglou (1989) propose an alternative estimator which uses the weight

$$\text{(3.7)} \qquad w_i(S) = \begin{cases} 1, & \text{if } \hat{N}_i \ge N_i \\ (\hat{N}_i/N_i)^{h-1}, & \text{otherwise,} \end{cases}$$

where h is subjectively chosen. They, however, suggest $h = 2$ as a general-purpose value. Note that the weights (3.6) and (3.7) are identical if one chooses $\delta = 1$ and $h = 2$.

To study the nature of the weights $w_i(D)$ or $w_i(S)$, let us consider the special case of simple random sampling of n elements from a population of N elements. In this case, $\hat{N}_i = N(n_i/n)$, where the random variable n_i is the sample size in ith domain. Taking $\delta = 1$ in (3.6), it now follows that $w_i(D) = w_i(S) = 1$ if n_i is at least as large as the expected sample size $E(n_i) = n(N_i/N)$, that is, the sample size dependent estimators can fail to borrow strength from related domains even when $E(n_i)$ is not large enough to make the direct estimator $\hat{Y}_i$ reliable. On the other hand, when $\hat{N}_i < N_i$ the weight $w_i(D)$, which equals $w_i(S)$ when $h = 2$, decreases as n_i decreases. As a

result, more weight is given to the synthetic component as n_i decreases. Thus, the weights behave well unlike in the case $\hat{N}_i \geq N_i$. Another disadvantage is that the weights do not take account of the size of between area variation relative to within area variation for the characteristic of interest, that is, all characteristics get the same weight irrespective of their differences with respect to between area homogeneity.

Holt, Smith and Tomberlin (1979) obtain a best linear unbiased prediction (BLUP) estimator of Y_i under the following model for the finite population:

$$(3.8)\qquad y_{ig\ell} = \mu_g + e_{ig\ell},\quad \ell = 1,\ldots,N_{ig};\quad g = 1,\ldots,G;\quad i = 1,\ldots,m$$

where $y_{ig\ell}$ is the y-value of the ℓth unit in the cell (i,g), μ_g's are fixed effects and the errors $e_{ig\ell}$ are uncorrelated with zero means and variances σ_g^2. Further, N_{ig} denotes the number of population elements in the large domain g that belong to the small area i. Suppose n_{ig} elements in a sample of size n fall in cell (i,g), and let $\bar{y}_{ig}$ and $\bar{y}_{\cdot g}$ denote the sample means for (i,g) and g, respectively.

The best linear unbiased estimator of μ_g under (3.8) is $\hat{\mu}_g = \bar{y}_{\cdot g}$ which in turn leads to the BLUP estimator of Y_i given by

$$\hat{Y}_i^B = \sum_g \hat{Y}_{ig}^C,$$

where $\hat{Y}_{ig}^C$ is a composite estimator of the total Y_{ig} giving the weight $w_{ig} = n_{ig}/N_{ig}$ to the direct estimator $\hat{Y}_{ig} = N_{ig}\bar{y}_{ig}$, and the weight $1 - w_{ig}$ to the synthetic estimator $\hat{Y}_{ig}^S = N_{ig}\bar{y}_{ig}$. It therefore follows that the BLUP estimator of Y_i tends to the synthetic estimator $\hat{Y}_{\cdot g}^S = \Sigma_g N_{ig}\bar{y}_{\cdot g}$ if the sampling fraction n_{ig}/N_{ig} is negligible for all g, irrespective of the size of between area variation relative to within area variation. This limitation of model (3.8) can be avoided by using more realistic models that include random area-specific effects. We consider such models in Section 4, and we obtain small area estimators under these models in Section 5 using a general EB or a variance components approach as well as a HB procedure.

4. SMALL AREA MODELS

We now consider small area models that include random area-specific effects. Two types of models have been proposed in the literature. In the first type, only area-specific auxiliary data $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^T$ are available and the parameters of interest, θ_i, are assumed to be related to $\mathbf{x}_i$. In particular, we assume that

$$(4.1)\qquad \theta_i = \mathbf{x}_i^T\beta + v_i z_i,\qquad i = 1,\ldots,m,$$

where the z_i's are known positive constants, β is the vector of regression parameters and the v_i's are independent and identically distributed (iid) random variables with

$$E(v_i) = 0,\qquad V(v_i) = \sigma_v^2.$$

In addition, normality of the random effects v_i is often assumed. In the second type of models, element-specific auxiliary data $\mathbf{x}_{ij} = (x_{ij1},\ldots,x_{ijp})^T$ are available for the population elements, and the variable of interest, y_{ij}, is assumed to be related to $\mathbf{x}_{ij}$ through a nested error regression model:

$$(4.2)\qquad y_{ij} = \mathbf{x}_{ij}^T\beta + v_i + e_{ij},\quad j = 1,\ldots,N_i;\quad i = 1,\ldots,m.$$

Here $e_{ij} = \tilde{e}_{ij}k_{ij}$ and the $\tilde{e}_{ij}$'s are iid random variables, independent of the v_i's, with

$$E(\tilde{e}_{ij}) = 0,\qquad V(\tilde{e}_{ij}) = \sigma^2,$$

the k_{ij}'s being known constants and N_i the number of elements in the ith area. In addition, normality of the v_i's and $\tilde{e}_{ij}$'s is often assumed. The parameters of inferential interest here are the small area totals Y_i or the means $\bar{Y}_i = Y_i/N_i$.

For making inferences about the θ_i's under model (4.1), we assume that direct estimators, $\hat{\theta}_i$, are available and that

$$(4.3)\qquad \hat{\theta}_i = \theta_i + e_i,\qquad i = 1,\ldots,m$$

where the e_i's are sampling errors, $E(e_i|\theta_i) = 0$ and $V(e_i|\theta_i) = \psi_i$, that is, the estimators $\hat{\theta}_i$ are design-unbiased. It is also customary to assume that the sampling variances, ψ_i, are known. These assumptions may be quite restrictive in some applications. For example, in the case of adjustment for census underenumeration, the estimates $\hat{\theta}_i$ obtained from a post-enumeration survey (PES) could be seriously biased, as noted by Freedman and Navidi (1986). Simlarly, if θ_i is a nonlinear function of the small area total Y_i and the sample size, n_i is small, then $\hat{\theta}_i$ may be seriously biased even if the direct estimator of Y_i is unbiased. We also assume normality of the $\hat{\theta}_i$'s, but this may not be as restrictive as the normality of the random effects v_i, due to the central limit theorem's effect on the $\hat{\theta}_i$'s.

Combining (4.3) and (4.1), we obtain the model

$$\hat{\theta}_i = \mathbf{x}_i^T\beta + v_i z_i + e_i, \qquad i = 1, \ldots, m \tag{4.4}$$

which is a special case of the general mixed linear model. Note that (4.4) involves design-induced random variables, e_i, as well as model-based random variables v_i.

Turning to the nested error regression model (4.2), we assume that a sample of size n_i is taken from the ith area and that selection bias is absent; that is, the sample values also obey the assumed model. The latter is satisfied under simple random sampling. It may also be noted that model (4.2) may not be appropriate under more complex sampling designs, such as stratified multistage sampling, since the design features are not incorporated. However, it is possible to extend this model to account for such features (see Section 7).

Writing model (4.2) in matrix form as

$$\mathbf{y}_i^P = \mathbf{X}_i^P\beta + v_i\mathbf{1}_i^P + \mathbf{e}_i^P, \tag{4.5}$$

where $\mathbf{X}_i^P$ is $N_i \times p$, $\mathbf{y}_i^P, \mathbf{e}_i^P$ and $\mathbf{1}_i^P$ are $N_i \times 1$ and $\mathbf{1}_i^P = (1, \ldots, 1)^T$, we can partition (4.5) as

$$\mathbf{y}_i^P = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{y}_i^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{X}_i^* \end{bmatrix}\beta + v_i \begin{bmatrix} \mathbf{1}_i \\ \mathbf{1}_i^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_i^* \end{bmatrix}, \tag{4.6}$$

where the superscript $*$ denotes the nonsampled elements. Now, writing the mean $\bar{Y}_i$ as

$$\bar{Y}_i = f_i\bar{y}_i + (1 - f_i)\bar{y}_i^*, \tag{4.7}$$

with $f_i = n_i/N_i$ and $\bar{y}_i$, $\bar{y}_i^*$ denoting the means for sampled and nonsampled elements respectively, we may view estimation of $\bar{Y}_i$ as equivalent to prediction of $\bar{y}_i^*$ given the data $\{\mathbf{y}_i\}$ and $\{\mathbf{X}_i\}$.

Various extensions of models (4.4) and (4.6), as well as models for binary and Poisson data, have been proposed in the literature. Some of these extensions will be briefly discussed in Section 7.

In the examples given in the Introduction, the models considered are special cases of (4.4) or (4.6). In Example 3, Ericksen and Kadane (1985, 1987) use model (4.4) with $z_i = 1$ and assume σ_v^2 to be known. Here $\hat{\theta}_i$ is a PES estimate of census undercount $\theta_i = \{(T_i - C_i)/T_i\}100$, where T_i is the true (unknown) count and C_i is the census count in the ith area. Cressie (1992) uses (4.4) with $z_i = C_i^{-1/2}$, where $\hat{\theta}_i$ is a PES estimate of the adjustment factor $\theta_i = T_i/C_i$. In Example 2, Fay and Herriot (1979) use (4.4) with $z_i = 1$, where $\hat{\theta}_i$ is a direct estimator of $\theta_i = \log P_i$ and P_i is the average percapita income (PCI) in the ith area. Further, $\mathbf{x}_i^T\beta = \beta_0 + \beta_1 x_i$ with x_i denoting the associated county value of log (PCI) from the 1970 census. In Example 4, Battese, Harter and Fuller (1988) use model (4.6) with $k_{ij} = 1$ and $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}$, where y_{ij}, x_{1ij} and x_{2ij} respectively denote the number of hectares of corn (or soybeans), the number of pixels classified as corn and the number of pixels classified as soybeans in the jth area segment of the ith county. A suitable model for our final example is also a special case of (4.6) with $\mathbf{x}_i^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$, where y_{ij} and x_{ij} respectively denote the total wages and salaries and gross business income for the jth firm in the ith area (census division).

5. EBLUP, EB AND HB APPROACHES

We now present the EBLUP, EB and HB approaches to small area estimation in the context of models (4.4) and (4.6). Both point estimation and measurement of uncertainty associated with the estimators will be studied.

5.1 EBLUP (Variance Components) Approach

As noted in Section 4, most small area models are special cases of a general mixed linear model involving fixed and random effects, and small area parameters can be expressed as linear combinations of these effects. Henderson (1950) derives BLUP estimators of such parameters in the classical frequentist framework. These estimators minimize the mean squared error among the class of linear unbiased estimators and do not depend on normality, similar to the best linear unbiased estimators (BLUEs) of fixed parameters. Robinson (1991) gives an excellent account of BLUP theory and examples of its application.

Under model (4.4), the BLUP estimator of $\theta_i = \mathbf{x}_i^T\beta + v_i z_i$ simplifies to a weighted average of the direct estimator $\hat{\theta}_i$ and the regression-synthetic estimator $\mathbf{x}_i^T\tilde{\beta}$:

$$\tilde{\theta}_i^H = \gamma_i\hat{\theta}_i + (1 - \gamma_i)\mathbf{x}_i^T\tilde{\beta}, \tag{5.1}$$

where the superscript H stands for Henderson,

$$\tilde{\beta} = \left[\sum_i \mathbf{x}_i\mathbf{x}_i^T/(\sigma_v^2 z_i^2 + \psi_i)\right]^{-1} \cdot \left[\sum_{i=1}^m \mathbf{x}_i\hat{\theta}_i/(\sigma_v^2 z_i^2 + \psi_i)\right] \tag{5.2}$$

is the BLUE estimator of β and

$$\gamma_i = \sigma_v^2 z_i^2/(\sigma_v^2 z_i^2 + \psi_i).$$

The weight γ_i measures the uncertainty in modelling the θ_is, namely, $\sigma_v^2 z_i^2$ relative to the total variance $\sigma_v^2 z_i^2 + \psi_i$. Thus, the BLUP estimator takes proper account of between area variation relative to the precision of the direct estimator. It is valid for general sampling designs since we are modelling only the θ_i's and not the individual elements in the population. It is also design consistent since $\gamma_i \to 1$ as the sampling variance $\psi_i \to 0$.

The mean squared error (MSE) of $\tilde{\theta}_i^H$ under model (4.4) may be written as

$$M_{1i}(\sigma_v^2) = E(\tilde{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2),$$

where

$$g_{1i}(\sigma_v^2) = \sigma_v^2 z_i^2 \psi_i (\sigma_v^2 z_i^2 + \psi_i)^{-1} = \gamma_i \psi_i$$

and

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \mathbf{x}_i^T \left[\sum_i \mathbf{x}_i \mathbf{x}_i^T / (\sigma_v^2 z_i^2 + \psi_i)\right]^{-1} \mathbf{x}_i.$$

The first term $g_{1i}(\sigma_v^2)$ is of order $O(1)$ while the second term $g_{2i}(\sigma_v^2)$, due to estimating β, is of order $O(m^{-1})$ for large m.

The BLUP estimator (5.1) depends on the variance component σ_v^2 which is unknown in practical applications. However, various methods of estimating variance components in a general mixed linear model are available, including the method of fitting constants or moments, maximum likelihood (ML) and restricted maximum likelihood (REML). Cressie (1992) gives a succinct account of these methods in the context of model (4.4). All these methods yield asymptotically consistent estimators under realistic regularity conditions.

Replacing σ_v^2 with an asymptotically consistent estimator $\hat{\sigma}_v^2$, we obtain a two-stage estimator, $\hat{\theta}_i^H$, which is referred to as the empirical BLUP or EBLUP estimator (Harville, 1991), in analogy with the EB estimator. It remains unbiased provided (i) the distributions of v_i and e_i are both symmetric (not necessarily normal); (ii) $\hat{\sigma}_v^2$ is an even function of $\hat{\theta}_i$'s and remains invariant when $\hat{\theta}_i$ is changed to $\hat{\theta}_i - \mathbf{x}_i^T \mathbf{a}$ for all $\mathbf{a}$ (Kackar and Harville, 1984). Standard methods of estimating variance components all satisfy (ii). We may also point out that the MSE of the EBLUP estimator appears to be insensitive to the choice of the estimator $\hat{\sigma}_v^2$.

If normality of the errors v_i also holds, then we can write the MSE of $\hat{\theta}_i^H$ as

$$M_{2i}(\sigma_v^2) = M_{1i}(\sigma_v^2) + E(\hat{\theta}_i^H - \tilde{\theta}_i^H)^2, \tag{5.3}$$

see Kackar and Harville (1984). It follows from (5.3) that the MSE of $\hat{\theta}_i^H$ is always larger than that of the BLUP estimator $\tilde{\theta}_i^H$. The second term of (5.3) is not tractable, unlike the first term $M_{1i}(\sigma_v^2)$; but it can be approximated for large m (Kackar and Harville, 1984; Prasad and Rao, 1990; Cressie, 1992). We have, for large m,

$$E(\hat{\theta}_i^H - \tilde{\theta}_i^H)^2 \doteq g_{3i}(\sigma_v^2) \tag{5.4}$$

where

$$g_{3i}(\sigma_v^2) = \psi_i^2 z_i^4 (\sigma_v^2 z_i^2 + \psi_i)^{-3} \overline{V}(\hat{\sigma}_v^2),$$

and the neglected terms in the approximation (5.4) are of lower order than $O(m^{-1})$. Here $\overline{V}(\hat{\sigma}_v^2)$ denotes the asymptotic variance of $\hat{\sigma}_v^2$; Cressie (1992) gives the asymptotic variance formulae for ML and REML estimators. It is customary to ignore the uncertainty in $\hat{\sigma}_v^2$ and use $M_{1i}(\hat{\sigma}_v^2) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2)$ as an estimator of MSE of $\hat{\theta}_i^H$, but this procedure could lead to severe underestimation of the true MSE. A correct, approximately unbiased estimator of MSE $(\hat{\theta}_i^H)$ is given by

$$\text{mse}(\hat{\theta}_i^H) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \tag{5.5}$$

(see Prasad and Rao, 1990). The bias of (5.5) is of lower order than m^{-1}.

Noting that $E[\Sigma_i (y_i - \mathbf{x}_i^T \tilde{\beta})^2 / (\sigma_v^2 z_i + \psi_i)] = m - p$, a method of moments estimator $\hat{\sigma}_v^2$ can be obtained by solving iteratively

$$\sum_{i=1}^{m} (y_i - x_i^T \tilde{\beta})^2 / (\sigma_v^2 z_i + \psi_i) = m - p$$

in conjunction with (5.2) and letting $\hat{\sigma}_v^2 = 0$ when no positive solution exists (Fay and Herriot, 1979). This method does not require normality, unlike the ML and REML. Alternatively, a simple moment estimator is given by $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$, where

$$\tilde{\sigma}_v^2 = (t-p)^{-1} \left[\sum_i \frac{1}{z_i^2} (y_i - \mathbf{x}_i^T \beta^*)^2 - \sum_i \frac{\psi_i}{z_i^2} \left\{1 - \mathbf{x}_i^T \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_i\right\}\right] \tag{5.6}$$

and $\beta^* = (\Sigma_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\Sigma_i \mathbf{x}_i \hat{\theta}_i)$ is the ordinary least squares estimator of β. The estimator $\tilde{\sigma}_v^2$ is unbiased for σ_v^2 and under normality,

$$\overline{V}(\hat{\sigma}_v^2) = \overline{V}(\tilde{\sigma}_v^2) = 2t^{-2} \sum_i (\sigma_v^2 + \psi_i / z_i^2)^2$$

(see Prasad and Rao, 1990 for the case $z_i = 1$).

Lahiri and Rao (1992) show that the estimator of MSE, (5.5), using the moment estimator (5.6), is also valid under moderate nonnormality of the random effects, v_i. Thus, inference based on $\hat{\theta}_i^H$ and $\text{mse}(\hat{\theta}_i^H)$ is robust to nonnormality of the random effects.

We next turn to the nested error regression model (4.6). The BLUP estimator of $\bar{Y}_i$ in this case is obtained as follows: (i) using the model $\mathbf{y}_i = \mathbf{X}_i + v_i \mathbf{1}_{n_i} + \mathbf{e}_i$ for the sampled elements, obtain the BLUP estimator of $\bar{\mathbf{X}}_i^{*T}\beta + v_i$, where $\bar{\mathbf{X}}_i^*$ is the mean for nonsampled elements; (ii) substitute this estimator for $\bar{y}_i^*$ in (4.7). Thus the BLUP estimator of $\bar{Y}_i$ is given by

$$(5.7)\quad \tilde{\bar{Y}}_i^H = f_i \bar{y}_i + (1 - f_i)\left[\bar{X}_i^{*T}\tilde{\beta} + \gamma_i(\bar{y}_{iw} - \bar{\mathbf{x}}_{iw}^T\tilde{\beta})\right],$$

where $\tilde{\beta}$ is the BLUE of β,

$$\gamma_i = \sigma_v^2(\sigma_v^2 + \sigma^2/w_{i\cdot})^{-1}$$

with $w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$ and $w_{ij} = k_{ij}^{-2}$, and $\bar{y}_{iw}$ and $\bar{\mathbf{x}}_{iw}$ are the weighted means with weights w_{ij} (see Prasad and Rao, 1990, and Stukel, 1991). The BLUE $\tilde{\beta}$ is readily obtained by applying ordinary least squares to the transformed data $\{(y_{ij} - \gamma_i \bar{y}_{iw})/k_{ij}, (\mathbf{x}_{ij} - \gamma_i \bar{\mathbf{x}}_{iw})/k_{ij}\}$ (see Stukel, 1991, and Fuller and Battese, 1973). If the sample fraction f_i is negligible, we can write $\tilde{\bar{Y}}_i^H$ as a composite estimator of the form

$$(5.8)\quad \tilde{\bar{Y}}_i^H \doteq \gamma_i[\bar{y}_{iw} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{iw})^T\tilde{\beta}] + (1 - \gamma_i)\bar{X}_i^T\tilde{\beta},$$

where $\bar{\mathbf{X}}_i$ is the ith area population mean of $\mathbf{x}_{ij}$'s. It follows from (5.8) that the BLUP estimator is a weighted average of the "survey regression" estimator $\bar{y}_{iw} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{iw})^T\tilde{\beta}$ and the regression synthetic estimator $\bar{\mathbf{X}}_i^T\tilde{\beta}$. If $k_{ij} = 1$ for all (ij), then the survey regression estimator is approximately design-unbiased for $\bar{Y}_i$ under simple random sampling even if n_i is small. In the case of general k_{ij}'s, it is model-unbiased conditional on the realized local effect v_i, unlike the BLUP estimator which is conditionally biased.

An empirical BLUP estimator, $\hat{\bar{Y}}_i^H$, is obtained from (5.7) by replacing (σ_v^2, σ^2) with asymptotically consistent estimators $(\hat{\sigma}_v^2, \hat{\sigma}^2)$. Further, assuming normality of the errors an approximately unbiased estimator of MSE $(\hat{\bar{Y}}_i^H)$, similar to (5.5) under model (4.4), is given by

$$(5.9)\quad \text{mse}(\hat{\bar{Y}}_i^H) = (1 - f_i)^2\left[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)\right].$$

Here

$$g_{1i}(\sigma_v^2, \sigma^2) = \gamma_i(\sigma^2/w_{i\cdot}) + (1 - f_i)^2 N_i^{-2}\mathbf{k}_i^{*T}\mathbf{k}_i^*$$

with $\mathbf{k}_i^*$ denoting the vector of k_{ij}'s for nonsampled units in ith area, and

$$g_{2i}(\sigma_v^2, \sigma^2) = (\bar{\mathbf{x}}_i^* - \gamma_i \bar{\mathbf{x}}_{iw})^T\mathbf{A}^{-1}(\bar{\mathbf{x}}_i^* - \gamma_i \bar{\mathbf{x}}_{iw})\sigma^2$$

with

$$\mathbf{A} = \sum_{i=1}^{m}\left[\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \gamma_i w_{i\cdot}\bar{\mathbf{x}}_{iw}\bar{\mathbf{x}}_{iw}^T\right].$$

Further,

$$g_{3i}(\sigma_v^2, \sigma^2) = w_{i\cdot}^{-2}(\sigma_v^2 + \sigma^2/w_{i\cdot})^{-3}\left[\sigma^2\overline{V}(\hat{\sigma}_v^2) + \sigma_v^2\overline{V}(\hat{\sigma}_e^2) - 2\sigma^2\sigma_v^2\overline{\text{cov}}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)\right],$$

where $\overline{\text{cov}}$ denotes the asymptotic covariance (see Stukel, 1991 and Prasad and Rao, 1990).

For the ML and REML methods, the asymptotic covariance matrix of $(\hat{\sigma}_v^2, \hat{\sigma}^2)$ can be obtained from general theory (see, e.g., Cressie, 1992). Stukel (1991) and Fuller and Battese (1973) use the method of fitting constants which involves two ordinary least square fittings: first, we calculate the residual sum of squares, SSE(1), with ν_1 degrees of freedom by regressing through the origin the y-deviations $k_{ij}^{-1}(y_{ij} - \bar{y}_{iw})$ on the nonzero x-deviations $k_{ij}^{-1}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{iw})$ for these areas with $n_i > 1$. Second, we calculate the residual sum of squares SSE(2) by regressing y_{ij}/k_{ij} on $\mathbf{x}_{ij}/k_{ij}$. Then $\hat{\sigma}^2 = \nu_1^{-1}$ SSE(1) and $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ with

$$\hat{\sigma}_v^2 = \eta_{**}^{-1}[SSE(2) - (n - p)\hat{\sigma}^2],$$

where

$$\eta^* = \sum_i w_{i\cdot}(1 - w_{i\cdot}\bar{\mathbf{x}}_{iw}^T\mathbf{A}_1^{-1}\bar{\mathbf{x}}_{iw})$$

with

$$\mathbf{A}_1 = \sum_i\sum_j w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T.$$

The Appendix gives the variances and covariance of $\hat{\sigma}^2$ and $\hat{\sigma}_v^2$.

Again, ignoring the uncertainty in $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ and using $M_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ as an estimator of MSE $(\hat{\bar{Y}}_i^H)$ could lead to severe underestimation of the true MSE.

Limited simulation results (Prasad and Rao, 1990; Datta and Ghosh, 1991 and Hulting and

Harville, 1991) indicate that the estimator of MSE, mse $(\widehat{\bar{Y}}_i^H)$, given by (5.9), performs well even for moderate m (as small as 15), provided σ_v^2/σ^2 is not close to zero.

5.2 EB Approach

In the EB approach, the posterior distribution of the parameters of interest given the data is first obtained, assuming that the model parameters are known. The model parameters are estimated from the marginal distribution of the data, and inferences are then based on the estimated posterior distribution. Morris (1983) gives an excellent account of the EB approach and significant applications.

Under model (4.4) with normal errors, the posterior distribution of θ_i given $\hat{\theta}_i, \beta$ and σ_v^2 is normal with mean θ_i^B and variance $g_{1i}(\sigma_v^2) = \gamma_i\psi_i$, where

$$\theta_i^B = E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i^T\beta.$$

Under quadratic loss, θ_i^B is the Bayes estimator of θ_i. Noting that the $\hat{\theta}_i \sim N(\mathbf{x}_i^T\beta,\ \sigma_v^2 z_i^2 + \psi_i)$ are marginally independent, we can obtain the estimators $\hat{\sigma}_v^2$ and $\hat{\beta}$ as before using ML, REML or the method of moments. The estimated posterior distribution is $N(\hat{\theta}_i^{EB}, g_{1i}(\hat{\sigma}_v^2))$, where $\hat{\theta}_i^{EB}$ is identical to the EBLUP estimator $\hat{\theta}_i^H$. A naive EB approach uses $\hat{\theta}_i^{EB}$ as the estimator of θ_i and measures its uncertainty by the estimated posterior variance

$$V(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2) = g_{1i}(\hat{\sigma}_v^2). \tag{5.10}$$

This can lead to severe underestimation of the true posterior variance $V(\theta_i|\hat{\theta})$ (under a prior distribution on β and σ_v^2), although $\hat{\theta}_i^{EB} = E(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2)$ is approximately equal to the true posterior mean $E(\theta_i|\hat{\theta})$, where $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T$.

The above point is better understood when one writes

$$E(\theta_i|\hat{\theta}) = E_{\beta,\sigma_v^2}[E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)]$$

and

$$\begin{aligned} V(\theta_i|\hat{\theta}) &= E_{\beta,\sigma_v^2}[V(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)] \\ &\quad + V_{\beta,\sigma_v^2}[E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)], \end{aligned} \tag{5.11}$$

where E_{β,σ_v^2} and V_{β,σ_v^2} respectively denote the expectation and variance with respect to the posterior distribution of β and σ_v^2 given the data $\hat{\theta}$. It follows from (5.11) that (5.10) is a good approximation only to the first variance term on the right side of (5.11), but the second variance term is ignored in the naive EB approach, that is, it fails to take account of the uncertainty about the parameters β and σ_v^2. Note that the form of the prior distribution on β and σ^2 is not specified in the EB approach, unlike in the HB approach (Section 5.3).

Two methods of accounting for the underestimation of true posterior variance have been proposed in the literature. The first method is based on the bootstrap (Laird and Louis, 1987), while the second method uses an asymptotic approximation to the posterior variance $V(\theta_i|\hat{\theta})$ irrespective of the form of the prior on β and σ_v^2 (Kass and Steffey, 1989). In the bootstrap method, a large number, B, of independent bootstrap samples $\{\theta_1^*(b), \ldots, \theta_m^*(b);\ b = 1, \ldots, B\}$ are first drawn, where $\theta_i^*(b)$ is drawn from the estimated marginal distribution $N(\mathbf{x}_i^T\hat{\beta}, \hat{\sigma}_v^2 z_i^2 + \psi_i)$. Estimates $\beta^*(b)$ and $\sigma_v^{*2}(b)$ are then computed from the bootstrap data $\{\theta_i^*(b), \mathbf{x}_i,\ i = 1, \ldots, m\}$ for each b. The EB bootstrap estimator of θ_i is given by

$$\begin{aligned} \theta_i^*(\cdot) &= \frac{1}{B}\sum_{b=1}^{B} E[\theta_i|\theta^*(b), \beta^*(b), \sigma_v^{*2}(b)] \\ &= \frac{1}{B}\sum_{b=1}^{B} \theta_i^{*EB}(b), \end{aligned}$$

and its uncertainty is measured by

$$\begin{aligned} V_i^* &= \frac{1}{B}\sum_{b=1}^{B} V[\theta_i|\theta_i^*(b), \beta^*(b), \sigma_v^{*2}(b)] \\ &\quad + \frac{1}{B-1}\sum_{b=1}^{B}[\theta_i^{*EB}(b) - \theta_i^{*EB}(\cdot)]^2. \end{aligned} \tag{5.12}$$

The second term on the right side of (5.12) accounts for the underestimation. The EB bootstrap method looks promising, but further studies on its frequentist performance are needed.

In the Kass-Steffey method, $\hat{\theta}_i^{EB}$ is taken as the estimator of θ_i, but a positive correction term is added to the estimated posterior variance $V(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2)$ to account for the underestimation. This term depends on the observed information matrix and the partial derivatives of θ_i^B, evaluated at the ML estimates $\hat{\beta}$ and $\hat{\sigma}_v^2$. This method also looks promising, but its frequentist properties remain to be investigated. (Steffey and Kass, 1991 conjecture that the MSE of EB estimator is approximately equal to their approximation to the posterior variance.) Kass and Steffey (1989) also give an improved second-order approximation to the true posterior variance, $V(\theta_i|\hat{\theta})$.

Turning to the nested error regression model (4.6), the estimated posterior distribution of $\bar{Y}_i$ given the data $\mathbf{y}$ is normal with mean equal to the EBLUP $\widehat{\bar{Y}}_i^H$ and variance equal to $(1-f_i)^2 g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ which

is a severe underestimate of the true posterior variance $V(\overline{Y}_i|\mathbf{y})$. Again, the bootstrap and Kass-Steffey methods can be applied to account for the underestimation.

If one wishes to view the EB approach in the frequentist framework, a prior distribution on β and σ_v^2 cannot be entertained. In this case, MSE is a natural measure of uncertainty and any differences between the EB and EBLUP approaches disappear under the normality assumption. It may also be noted that the EB estimator can be justified without the normality assumption, similar to the EBLUP, using the "posterior linearity" property (Ghosh and Lahiri, 1987; Ericson, 1969).

5.3 HB Approach

In the HB approach, a prior distribution on the model parameters is specified and the posterior distribution of the parameters of interest is then obtained. Inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance. The HB approach is straightforward and clear-cut but computationally intensive, often involving high dimensional integration. Recent advances in computational aspects of the HB approach, such as Gibbs sampling (cf. Gelfand and Smith, 1990) and importance sampling, however, seem to overcome the computational difficulties to a large extent. If the solution involves only one or two dimensional integration, it is often easier to perform direct numerical integration than to use Gibbs sampling or any other Monte Carlo numerical integration method. Datta and Ghosh (1991) apply the HB approach to estimation of small area means, $\overline{Y}_i$, under general mixed linear models, and also discuss the computational aspects.

We now illustrate the HB approach under our models (4.4) and (4.6), assuming noninformative priors on β and the variance components σ_v^2 and σ^2. The HB approach, however, can incorporate prior information on these parameters through informative priors.

Under model (4.4), we first obtain the posterior distribution of θ_i given $\hat{\theta}$ and σ_v^2, by assuming that β has a uniform distribution over R^p to reflect absence of prior information on β. Straightforward calculations show that it is normal with mean equal to the BLUP estimator $\tilde{\theta}_i^H$ and variance equal to $M_{1i}(\sigma_v^2)$, the MSE of $\tilde{\theta}_i^H$, that is, $E(\theta_i|\hat{\theta},\sigma_v^2) = \tilde{\theta}_i^H$ and $V(\theta_i|\hat{\theta},\sigma_v^2) = \text{MSE}\,(\tilde{\theta}_i^H)$. Hence, when σ_v^2 is assumed to be known, the HB and BLUP approaches lead to identical inferences.

To take account of the uncertainty about σ_v^2, we need to calculate the posterior distribution of σ_v^2 given $\hat{\theta}$ under a suitable prior on σ_v^2. The posterior mean and variance of θ_i are then given by

$$E(\theta_i|\hat{\theta}) \equiv E_{\sigma_v^2}(\tilde{\theta}_i^H) \tag{5.13}$$

and

$$V(\theta_i|\hat{\theta}) = E_{\sigma_v^2}[M_{1i}(\sigma_v^2)] + V_{\sigma_v^2}(\tilde{\theta}_i^H), \tag{5.14}$$

where $E_{\sigma_v^2}$ and $V_{\sigma_v^2}$ respectively denote the expectation and variance with respect to the posterior distribution of σ_v^2 given $\hat{\theta}$. Numerical evaluation of (5.13) and (5.14) involves one dimensional integration. Ghosh (1992) obtains the posterior distribution, $f(\sigma_v^2|\hat{\theta})$, assuming that σ_v^2 has a uniform distribution over $(0,\infty)$ to reflect the absence of prior information about σ_v^2, and that σ_v^2 and β are independently distributed. It is given by

$$f(\sigma_v^2|\hat{\theta}) = (\sigma_v^2)^{-\frac{(m-p)}{2}} \left\{\prod_1^m \gamma_i^{1/2}\right\} \left|\sum_i \gamma_i \mathbf{x}_i \mathbf{x}_i^T\right|^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2} Q_a(\hat{\theta})\right],$$

where

$$Q_a(\hat{\theta}) = (\sigma_v^2)^{-1}\left[\sum_i \gamma_i \hat{\theta}_i^2 - \left(\sum_i \gamma_i \hat{\theta}_i \mathbf{x}_i\right)^T \cdot \left(\sum_i \gamma_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_i \gamma_i \hat{\theta}_i \mathbf{x}_i\right)\right].$$

We next turn to the nested error regression model (4.6). We first obtain the posterior distribution of $\overline{Y}_i$ given $\mathbf{y}$, σ_v^2 and σ^2, by assuming that β has uniform distribution over R^p. Straightforward calculations show that it is normal with mean equal to the BLUP estimator $\tilde{\overline{Y}}_i^H$ and variance equal to $\text{MSE}\,(\tilde{\overline{Y}}_i^H) = M_{1i}(\sigma_v^2,\sigma^2)$, that is, $E(\overline{Y}_i|\mathbf{y},\sigma_v^2,\sigma^2) = \tilde{\overline{Y}}_i^H$ and $V(\overline{Y}_i|\mathbf{y},\sigma_v^2,\sigma^2) = \text{MSE}(\tilde{\overline{Y}}_i^H)$. Hence, when both σ_v^2 and σ^2 are assumed to be known, the HB and BLUP approaches lead to identical inferences.

To take account of the uncertainty about σ_v^2 and σ^2, Datta and Ghosh (1991) further assume $\beta, (\sigma^2)^{-1}$ and $(\sigma_v^2)^{-1} = (\sigma^2)^{-1}\lambda$ to be independently distributed with $(\sigma^2)^{-1} \sim$ gamma $((1/2)a_0, (1/2)g_0)$ and $(\sigma^2)^{-1}\lambda \sim$ gamma $((1/2)a_1, (1/2)g_1)$, where $a_0 \geq 0$, $g_0 \geq 0$, $a_1 > 0$, $g_1 \geq 0$ and $\lambda = \sigma^2/\sigma_v^2$. Here gamma (α,β) denotes the gamma random variable with pdf $f(z) = \exp(-\alpha z)\alpha^\beta z^{\beta-1}/\Gamma(\beta)$, $z > 0$. Datta and Ghosh (1991) obtain closed form expressions for $E(\overline{Y}_i|\mathbf{y},\lambda)$

and $V(\bar{Y}_i|\mathbf{y},\lambda)$ by showing that $f(\mathbf{y}^*|\mathbf{y},\lambda)$ is a multivariate t-distribution. They also derive the posterior distribution of λ given $\mathbf{y}$, but it has a complex structure making it necessary to perform one-dimensional numerical integration to get $E(\bar{Y}_i|\mathbf{y})$ and $V(\bar{Y}_i|\mathbf{y})$ using the following relationships:

$$E(\bar{Y}_i|\mathbf{y}) = E_\lambda[E(\bar{Y}_i|\mathbf{y},\lambda)]$$

and

$$V(\bar{Y}_i|\mathbf{y}) = E_\lambda[V(\bar{Y}_i|\mathbf{y},\lambda)] + V_\lambda[E(\bar{Y}_i|\mathbf{y},\lambda)],$$

where E_λ and V_λ respectively denote the expectation and variance under the posterior distribution of λ given the data $\mathbf{y}$.

Datta and Ghosh (1991) compare the HB, EB and EBLUP approaches using the data for our example 4 and letting $a_0 = a_1 = 0.005$ and $g_0 = g_1 = 0$ to reflect the absence of prior information on σ_v^2 and σ^2. As one might expect, the three estimates were close to each other as point predictors of small area (county) means; the EB estimate was obtained by replacing λ with the method-of-fitting constants estimate $\hat{\lambda}$ in $E(\bar{Y}_i|\mathbf{y},\lambda)$. The naive variance estimate, $V(\bar{Y}_i|\mathbf{y},\hat{\lambda}) = (s_i^{EB})^2$ associated with the EB estimate $E(\bar{Y}_i|\mathbf{y},\hat{\lambda})$, was always found to be smaller than the true posterior variance, $V(\bar{Y}_i|\mathbf{y}) = (s_i^{HB})^2$, associated with the HB estimate $\hat{\bar{Y}}_i^{HB} = E(\bar{Y}_i|\mathbf{y})$; for one county, s_i^{EB} was about 10% smaller than s_i^{HB}. Note that the customary naive EB variance estimate, $V(\bar{Y}_i|\mathbf{y},\hat{\beta},\hat{\sigma}_v^2,\hat{\sigma}^2)$, will lead to much more severe underestimation than $V(\bar{Y}_i|\mathbf{y},\hat{\lambda})$ since the latter takes account of the uncertainty about β and σ^2. The estimated MSE, mse $(\hat{\bar{Y}}_i^H) = (s_i^H)^2$, associated with the EBLUP estimate, $\hat{\bar{Y}}_i^H$, was found to be similar to the HB variance estimate. Our example in Section 6 also gives similar results. Datta and Ghosh (1991) have also conducted a simulation study on the frequentist properties of the HB and EBLUP methods using the Battese, Harter and Fuller (1988) model. Their findings indicate that the simulated MSEs for the HB estimator are very close to those for the EBLUP estimator while the coverage probabilities based on $\hat{\bar{Y}}_i^{HB} \pm (1.96)s_i^{HB}$ turn out to be slightly bigger than those based on $\hat{\bar{Y}}_i^H \pm (1.96)s^H$, both being close to nominal confidence level of 95%. Hulting and Harville (1991) obtain similar results in another simulation study using the Battese, Harter and Fuller (1988) model and varying the variance ratio σ_v^2/σ^2. However, they find the HB method produces different and more sensible answers than the EBLUP procedure if the estimate for σ_v^2/σ^2 is zero or close to zero.

The HB approach looks promising, but we need to study its robustness to choice of prior distributions on the model parameters.

6. EXAMPLE

Several of the proposed small area estimators are now compared on the basis of their squared errors and relative errors from the true small area means $\bar{Y}_i$. For this purpose, we first constructed a synthetic population of pairs (y_{ij}, x_{ij}) resembling the business population studied by Särndal and Hidiroglou (1989) where the census divisions are small areas, y_{ij} denotes wages and salaries of jth firm in the ith census division and x_{ij} the corresponding gross business income. To generate the synthetic population, we fitted the nested error regression model (4.6) with $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$ to a real population to estimate β_0 and β_1 and the variance components σ_v^2 and σ^2. The resulting synthetic model is given by

$$\begin{aligned} y_{ij} &= -2.47 + 0.20x_{ij} + v_i + e_{ij}, \\ j &= 1,\ldots,N_i,\ i = 1,\ldots,m, \\ v_i &\overset{\text{i.i.d}}{\sim} N(0, 22.14), \\ e_{ij} &\overset{\text{i.i.d}}{\sim} N(0, 0.47x_{ij}). \end{aligned} \tag{6.1}$$

We then used model (6.1) in conjunction with the population x_{ij}-values to generate a synthetic population of pairs (y_{ij}, x_{ij}) with $m = 16$ small areas. Table 1 reports the small area population sizes, N_i, and the small area means $(\bar{Y}_i, \bar{X}_i)$ for this synthetic population of size $N = 114$. A simple random sample of size $n = 38$ was drawn from the synthetic population. The resulting small area sample sizes, n_i, and sample data (y_{ij}, x_{ij}) are reported in Table 2. Note that direct estimators cannot be implemented for areas 1, 4 and 13 since $n_i = 0$ for these areas. We have, therefore, confined ourselves to the following indirect estimators valid for all $n_i \geq 0$:

(i) Ratio-synthetic estimator: $\hat{\bar{Y}}_i^{RS} = (\bar{y}/\bar{x})\bar{X}_i$, where $(\bar{y}, \bar{x})$ are the overall sample means.

(ii) Sample-size dependent estimator:

$$\hat{\bar{Y}}_i^{SD} = \begin{cases} \hat{\bar{Y}}_i^{\text{REG}} = \bar{y}_i + (\bar{y}/\bar{x})(\bar{X}_i - \bar{x}_i), & \text{if } w_i \geq W_i, \\ \frac{w_i}{W_i}(\hat{\bar{Y}}_i^{\text{REG}}) + \left(1 - \frac{w_i}{W_i}\right)\hat{\bar{Y}}_i^{RS}, & \text{if } w_i < W_i, \end{cases}$$

where $\hat{\bar{Y}}_i^{\text{REG}}$ is a "survey regression" estimator, $(\bar{y}_i, \bar{x}_i)$ are the sample means, $w_i = n_i/n$ and $W_i = N_i/N$. This estimator corresponds to the weight (3.6) with $\delta = 1$ or the weight

TABLE 1
Small area sizes, N_i, and means $(\bar{Y}_i, \bar{X}_i)$ for a synthetic population ($N = 114$)

Area No.	N_i	$\overline{X}_i$	$\overline{Y}_i$	Area No.	N_i	$\overline{X}_i$	$\overline{Y}_i$
1	1	137.70	24.22	9	27	97.58	15.56
2	6	100.84	20.43	10	5	76.04	5.88
3	4	47.72	5.48	11	12	90.15	15.20
4	1	45.64	6.55	12	7	86.24	13.40
5	8	108.53	20.55	13	4	164.28	26.06
6	6	65.68	14.85	14	6	164.70	22.44
7	6	116.34	21.46	15	13	83.86	9.40
8	6	92.74	13.40	16	2	134.49	29.49

TABLE 2
Data from a simple random sample drawn from a synthetic population ($n = 38, N = 114$)

Area No.	n_i	x_{ij}	y_{ij}	Area No.	n_i	x_{ij}	x_{ij}
1	0	—	—	9	10	333.24	47.62
						80.91	5.27
2	3	33,70	5.90			43.65	6.97
		47.19	13.22			29.29	−0.19
		75.21	17.44			102.66	15.94
						109.34	19.84
3	1	36.43	2.54			30.56	2.57
						127.96	24.61
4	0	—	—			190.34	35.41
						52.16	2.54
5	1	28.82	3.61	10	1	45.91	−6.34
				11	2	43.03	8.83
6	2	30.60	11.48			190.12	27.31
		129.69	21.45	12	1	47.39	1.70
				13	0	—	—
7	4	200.60	46.96	14	3	35.66	−0.80
		113.92	15.57			40.23	2.75
		74.33	8.66			111.23	10.87
		53.00	11.90	15	6	51.61	−3.20
						67.46	12.47
8	3	95.43	11.76			190.97	21.77
		35.75	−0.69			35.11	2.92
		39.08	21.46			25.09	−5.46
						73.51	7.35
				16	1	229.32	53.83

(3.7) with $h = 2$. We have not included the optimal composite estimator due to difficulties in estimating the optimal weight (3.4).

(iii) EBLUP (or EB) estimator $\hat{\overline{Y}}_i^H$ under model (4.6) with $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$, where σ_v^2 and σ^2 are estimated by the method of fitting constants.

(iv) HB estimator $\hat{\overline{Y}}_i^{HB}$ under model (4.6) as in (iii), using Datta-Ghosh's diffuse priors with $a_0 = 0$, $g_0 = 0$, $a_1 = 0.05$ and $g_1 = 0$.

Using the sample data (y_{ij}, x_{ij}) and the known small area population means $\overline{X}_i$ we computed the above four estimates along with their average relative errors

$$\text{ARE} = \frac{1}{m}\sum_{i=1}^{m} |\text{est.} - \overline{Y}_i|/\overline{Y}_i$$

and average squared errors

$$\text{ASE} = \frac{1}{m}\sum_{i=1}^{m} (\text{est.} - \overline{Y}_i)^2.$$

These values are reported in Table 3. We also calculated the standard error, s_i^H, of EBLUP estimator using (5.9) and the posterior standard deviation

TABLE 3
Small area estimates and their (%) average relative errors and average squared roots; standard error (S.E.) of EBLUP and HB estimators

Area No.	n_i	$\bar{Y}_i$	RS	SD	EBLUP	HB	S.E. EBLUP	S.E. HB
1	0	24.22	19.79	19.79	22.16	22.16	7.40	8.29
2	3	20.43	14.90	19.20	20.47	20.18	2.20	2.47
3	1	5.48	6.86	5.34	4.85	4.87	2.62	2.60
4	0	6.55	6.56	6.56	4.97	4.94	5.40	5.99
5	1	20.55	15.60	15.52	17.98	17.81	3.10	3.17
6	2	14.85	9.44	14.39	13.99	13.47	2.07	2.40
7	4	21.46	16.72	21.62	21.31	21.22	1.59	1.74
8	3	13.40	13.33	11.22	11.44	11.58	1.86	2.00
9	10	15.56	14.02	14.27	13.95	13.98	1.14	1.22
10	1	5.88	10.93	6.27	3.30	3.96	3.06	3.63
11	2	15.20	12.96	13.29	14.66	14.44	2.61	2.57
12	1	13.40	12.11	11.17	9.97	10.17	3.14	3.14
13	0	26.06	23.61	23.61	27.13	27.13	5.52	6.13
14	3	22.44	23.67	18.98	24.05	24.22	3.10	3.48
15	6	9.40	12.05	7.40	8.24	8.43	1.32	1.50
16	1	29.49	19.33	40.20	30.31	30.24	2.58	2.87
Av. Rel. Error%:			17.85	12.40	11.74	11.23		
Av. Sq. Error:			22.10	12.38	2.84	2.69		

RS=ratio synthetic estimator; SD=sample-size dependent estimator; EBLUP=EBLUP or EB estimator; HB=HB estimator.

(standard error), s_i^{HB}, of HB estimator using one-dimensional numerical integration. These values are also reported in Table 3.

The following observations on the relative performances of small area estimates may be drawn from Table 3: (1) EBLUP and HB estimators give similar values over small areas, and their average relative errors (%) are 11.74 and 11.23 and squared errors are 2.84 and 2.69 respectively. Asymptotically (as $m \to \infty$), the two estimators are identical, and the observed differences are due to moderate $m(= 16)$ and the method of estimating σ_v^2 and σ^2 (REML or ML would give slightly different EBLUP values). (2) Standard error values for EBLUP and HB estimators are also similar. This is in agreement with the empirical results of Datta and Ghosh (1991) and Hulting and Harville (1991). (3) Under the criterion of average squared error, EBLUP and HB estimators perform much better than the ratio-synthetic and sample-size dependent estimators: 2.84 for EBLUP vs. 12.38 for sample-size dependent (SD) and 22.10 for ratio-synthetic (RS). (4) Under the criterion of average relative error (%), however, EBLUP and HB estimates are not much better than the sample-size dependent estimator: 11.74 for EBLUP versus 12.40 for SD. However, both perform much better than the ratio-synthetic estimator with % ARE = 17.85.

It may be noted that EBLUP, EB and HB estimators are optimal under squared error loss and cease to be so under relative error loss. This is due to the fact that the Bayes estimators under relative error loss can often differ quite significantly from those under squared error loss. This nonoptimality carries over to EBLUP estimator which usually mimics closely the Bayes estimators. The above observations could perhaps explain why in our example the Bayes and EBLUP estimator did not improve significantly over the SD estimator under relative error.

All in all, our results in Table 3 clearly demonstrate the advantages of using the EBLUP or HB estimator and associated standard error when the assumed random effects model fits the data well. (Note that we simulated the data from an assumed model.) It is important, therefore, to examine the aptness of the assumed model using suitable diagnostic tools; Section 7.1 gives a brief account of diagnostics for models (4.4) and (4.6).

7. SPECIAL PROBLEMS

In this section we focus on special problems that may be encountered in implementing model-based methods for small area estimation. We also discuss some extensions of our basic models (4.4) and (4.6).

7.1 Model Diagnostics

Model-based methods rely on careful checking of the assumed models in order to find suitable models

that fit the data well. Model diagnostics, therefore, play an important role. However, the literature on diagnostics for mixed linear models involving random effects is not extensive, unlike standard regression diagnostics. Only recently have some useful diagnostic tools been proposed. See, for example, Battese, Harter and Fuller (1988); Beckman, Nachtsheim and Cook (1987); Calvin and Sedransk (1991); Christensen, Pearson and Johnson (1992); Cressie (1992); Dempster and Ryan (1985) and Lange and Ryan (1989).

We first consider the Fay-Herriot type model (4.4), where only area-specific covariates are used. When the model is correct, the standardized residuals $r_i = (\hat{\sigma}_v^2 z_i^2 + \psi_i)^{-(1/2)} (\hat{\theta}_i - \mathbf{x}_i^T \hat{\beta})$, $i = 1, \ldots, m$ are approximately iid $N(0,1)$ for large m where $\hat{\beta}$ is the BLUE estimator (5.2) with σ_v^2 replaced by $\hat{\sigma}_v^2$. We can, therefore, use a $q - q$ plot of r_i against $\Phi^{-1}[F_m(r_i)]$, where $\Phi(r)$ and $F_m(r)$ are the standard normal and empirical cdfs, respectively. A primary goal of this plot is to check the normality of the random effects v_i since the sampling errors e_i are approximately normal due to the central limit theorem effect. Dempster and Ryan (1985) note that the above $q - q$ plot may be inefficient for this purpose since it gives equal weight to each observation, even though the $\hat{\theta}_i$s differ in the amount of information contained about the v_is. They propose a weighted $q - q$ plot which uses a weighted empirical cdf $F_m^*(r) = \Sigma_i I(r - r_i) W_i / \Sigma_i W_i$ in place of $F_m(r)$, where $I(t) = 1$ for $t \geq 0$ and 0 otherwise, and $W_i = (\hat{\sigma}_v^2 + z_i^{-2} \psi_i)^{-1}$ in our case. This plot is more sensitive to departures from normality than the unweighed plot since it assigns greater weight to those observations for which $\hat{\sigma}_v^2$ account for a larger part of the total variance $\hat{\sigma}_v^2 + z_i^{-2} \psi_i$.

We next turn to the nested errors regression model (4.6), where the y_{ij}'s are correlated for each i. In this case, the transformed residuals $r_{ij} = k_{ij}^{-1} (y_{ij} - \hat{\gamma}_i \bar{y}_{iw}) - k_{ij}^{-1} (\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_{iw})^T \hat{\beta}$ are approximately uncorrelated with equal variances σ^2. Therefore, traditional regression diagnostics may be applied to the r_{ij}s, but the transformation can mask the effect of individual errors e_{ij}. On the other hand, standardized BLUP residuals $k_{ij}^{-1} (y_{ij} - \mathbf{x}_{ij}^T \hat{\beta} - \hat{v}_i) / \hat{\sigma}$ may be used to study the effect of individual units (ij) on the model, provided they are not strongly correlated. Lange and Ryan (1989) propose methods for checking the normality assumption on the random effects v_i using the BLUP estimates $\hat{v}_i$.

Christensen, Pearson and Johnson (1992) develop case-deletion diagnostics for detecting influential observations in mixed linear models. Their methods can be applied to model (4.6) as well as to more complex small area models.

7.2 Constrained Estimation

Direct survey estimates are often adequate at an aggregate (or large area) level in terms of precision. For example, Battese, Harter and Fuller (1988), in their application, find that the direct regression estimator of the mean crop area for the 12 counties together has adequate precision. It is, therefore, sometimes desirable to modify the individual small area estimators so that a properly weighted sum of these estimators equals the model-free, direct estimator at the aggregate level. The modified estimators will be somewhat less efficient than the original, optimal estimators, but they avoid possible aggregation bias by ensuring consistency with the direct estimator. One simple way to achieve consistency is to make a ratio adjustment, for example, the EBLUP estimator $\hat{Y}_i^H$ of a total Y_i is modified to

$$\hat{Y}_i^H \text{ (mod)} = \left(\hat{Y}_i^H / \sum_i \hat{Y}_i^H \right) \hat{Y}, \tag{7.1}$$

where $\hat{Y}$ is a direct estimator of the aggregate population total $Y = \Sigma_i Y_i$. Battese, Harter and Fuller (1988) and Pfeffermann and Barnard (1991) propose alternative estimators involving estimated variances and covariances of the optimal estimators $\hat{Y}_i^H$.

The previous sections focused on simultaneous estimation of small area means or totals, but in some applications the main objective is to produce an ensemble of parameter estimates whose histogram is in some sense close to the histogram of small area parameters. Spjøtvoll and Thomsen (1987), for example, were interested in finding how 100 municipalities in Norway were distributed according to proportion of persons not in the labor force. They propose constrained *EB* estimators whose variation matched the variation of the small area population means. By comparing with the actual distribution in their example, they show that the *EB* estimators are biased toward the prior mean compared to the constrained *EB* estimators. Constrained estimators reduce shrinking towards the synthetic component; for example, in (5.1) the weight $1 - \gamma_i$, attached to the synthetic component, is reduced to $1 - \gamma_i^{1/2}$. Following Louis (1984), Ghosh (1992) develops a general theory of constrained HB estimation. Ghosh obtains constrained HB estimates by matching the first two moments of the histogram of the estimates, and the posterior expectations of the first two moments of the histogram of the parameters and minimizing, subject to these conditions, the posterior expectation of the Euclidean distance between the estimates and the parameters. Lahiri (1990) obtains

similar results in the context of small area estimation, assuming "posterior linearity," thus avoiding distributional assumptions. Constrained Bayes estimates are suitable for subgroup analysis where the problem is not only to estimate the different components of a parameter vector but also to identify the parameters that are above or below a specified cut-off point. It should be noted that synthetic estimates are inappropriate for this purpose.

The optimal estimators (i.e., EBLUP, EB and HB estimators) may perform well overall but poorly for particular small areas that are not consistent with the assumed model on small area effects. To avoid this problem, Efron and Morris (1972) and Fay and Harriot (1979) suggest a straightforward compromise that consists of restricting the amount by which the optimal estimator differs from the direct estimator by some multiple of the standard error of the direct estimator. For example, a compromise estimator corresponding to the HB estimator $\hat{\theta}_i^{HB}$, under a normal prior on the θ_i's, is given by

$$\tilde{\theta}_i^{HB} = \begin{cases} \hat{\theta}_i^{HB}, & \text{if } \hat{\theta}_i - c\psi_i^{1/2} \le \hat{\theta}_i^{HB} \le \hat{\theta}_i + c\psi_i^{1/2} \\ \hat{\theta}_i - c\psi_i^{1/2}, & \text{if } \hat{\theta}_i^{HB} < \hat{\theta}_i - c\psi_i^{1/2} \\ \hat{\theta}_i + c\psi_i^{1/2}, & \text{if } \hat{\theta}_i^{HB} > \hat{\theta}_i + c\psi_i^{1/2}, \end{cases}$$

where $c > 0$ is a suitable chosen constant, say $c = 1$. A limitation of the compromise estimators is that no reliable measures of their precision are available.

7.3 Extensions

Various extensions of the basic models (4.4) and (4.6) have been studied in the literature. Due to space limitation, we can only mention some of these extensions.

Datta et al. (1992) extend the aggregate-level model (4.4) to the case of correlated sampling errors with a known covariance matrix and develop HB and EB estimators and associated measures of precision. In their application to adjustment of census undercount, the sampling covariance matrix is block diagonal. Cressie (1990a) introduces spatial dependence among the random effects v_i, in the context of adjustment for census undercount. Fay (1987) and Ghosh, Datta and Fay (1991) extend model (4.4) to multiple characteristics and perform hierarchical and empirical multivariate Bayes analysis, assuming that the sampling covariance matrix of $\hat{\theta}_i$, the vector of direct estimators for ith area, is known for each i. In their application to estimation of median income for four-person families by state, $\theta_i = (\theta_{i1}, \theta_{i2})^T$ with θ_{i1} = population median income of four-person families in state i and $\theta_{i2} = \frac{3}{4}$ (population median income of three-person families in state i) $+\frac{1}{4}$ (median income of four-person families in state i). By taking advantage of the strong correlation between the direct estimators $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$, they were able to obtain improved estimators of θ_{i1}.

Many surveys are repeated in time with partial replacement of the sample elements, for example, the monthly U.S. Current Population Survey and the Canadian Labor Force Survey. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. Cronkite (1987) developed regression synthetic estimators using pooled cross-sectional time series data and applied them to estimate substate area employment and unemployment using the Current Population Survey monthly survey estimates as dependent variable and counts from the Unemployment Insurance System and Census variables as independent variables. Rao and Yu (1992) propose an extension of model (4.4) to time series and cross-sectional data. Their model is of the form

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \qquad t = 1, \ldots, T, \tag{7.2}$$

$$\theta_{it} = \mathbf{x}_{it}'\beta + v_i + u_{it}, \qquad i = 1, \ldots, m, \tag{7.3}$$

where $\hat{\theta}_{it}$ is the direct estimator for small area i at time t, the e_{it}'s are sampling errors with a known block diagonal covariance matrix Ψ = block diag $(\Psi_1, \ldots, \Psi_m)$, $\mathbf{x}_{it}$ is a vector of covariates and $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$. Further, the u_{it}'s are assumed to follow a first order autoregressive process for each i, i.e., $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, $|\rho| < 1$ with $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. They obtain the EBLUP and HB estimators and their standard errors under (7.2) and (7.3).

Models of the form (7.3) have been extensively used in the econometric literature, ignoring sampling errors (see, e.g., Anderson and Hsiao, 1981; Judge, 1985, Chapter 13). Choudhry and Rao (1989) treat the composite error $w_{it} = e_{it} + u_{it}$ as a first order autoregressive process and obtain the EBLUP estimator of $\mathbf{x}_{it}^T\beta + v_i$. A drawback of their method is that the area by time specific effect u_{it} is ignored in modelling the θ_{it}'s.

Pfeffermann and Burck (1990) investigate more general models on the θ_{it}'s, but they assume modeling of sampling errors across time. They obtain EBLUP estimators of small area means using the Kalman filter. Singh and Mantel (1991) consider arbitrary covariance structures on sampling errors and propose recursive composite estimators using the Kalman filter. These estimators are not optimal but appear to be quite efficient relative to the corresponding EBLUP estimators.

Turning to extension of the nested error regression model (4.6), Fuller and Harter (1987) propose a multivariate nested error regression model and obtain EBLUP estimators and associated standard errors. Stukel (1991) studies two-fold nested error regression models, and obtains EBLUP estimators and associated standard errors. Such models are appropriate for two-stage sampling within small areas. Kleffe and Rao (1992) extend model (4.6) to the case of random error variances, σ_i^2, and obtain EBLUP estimator and associated standard errors in the special case of $\mathbf{x}_{ij} = 1$.

MacGibbon and Tomberlin (1989) and Malec, Sedransk and Tompkins (1991) study logistic regression models with random area-specific effects. Such models are appropriate for binary response variables when element-specific covariates are available. MacGibbon and Tomberlin (1989) obtain EB estimators of small area proportions and associated standard errors, but they ignore the uncertainty about the prior parameters. Farrell, MacGibbon and Tomberlin (1992) apply the bootstrap method of Laird and Louis (1987) to account for the underestimation of true posterior variance. Malec, Sedransk and Tompkins (1991) obtain HB estimators and associated standard errors using Gibbs sampling and apply their method to data from the U.S. National Health Interview Survey to produce estimates of proportions for individual states.

EB and HB methods have also been used for estimating regional mortality and disease rates (see, e.g., Marshall, 1991). In these applications, the observed small area counts, y_i, are assumed to be independent Poisson with conditional mean $E(y_i|\theta_i) = n_i\theta_i$, where θ_i and n_i respectively denote the true rate and number exposed in the ith area. Further, the θ_is are assumed to be random with a specified distribution (e.g., a gamma distribution with unknown scale and shape parameters). The EB or HB estimators are shrinkage estimators in the sense that the crude rate y_i/n_i is shrunk towards an overall regional rate, ignoring the spatial aspect of the problem. Marshall (1991) proposes "local" shrinkage estimators obtained by shrinking the crude rate towards a local neighbourhood rate. Such estimators are practically appealing and further work on their statistical properties is desirable.

De Souza (1992) studies joint mortality rates of two cancer sites over several geographical areas and obtains asymptotic approximations to posterior means and variances using the general first order approximations given by Kass and Steffey (1989). The bivariate model leads to improved estimators for each site compared to the estimators based on univariate models.

8. CONCLUSION

In this article, we have used the term "small area" to denote any local geographical area that is small or to describe any small subgroup of a population such as a specific age-sex-race group of people within a large geographical area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide desired accuracy at a much higher level of aggregation. As a result, the usual direct estimators of a small area mean are unlikely to give acceptable reliability; and it becomes necessary to "borrow strength" from related areas to find more accurate estimators for a given area or, simultaneously, for several areas. Considerable attention has been given to such indirect estimators in recent years.

We have attempted to provide an appraisal of indirect estimation covering both traditional design-based methods and newer model-based approaches to small area estimation. Traditional methods covered here include demographic techniques for local estimation of population and other characteristics of interest in post-censal years, and synthetic and sample size dependent estimation. Model-based methods studied here include EBLUP, EB and HB estimation. Two types of basic small area models that include random area-specific effects are used to describe these methods. In the first type of models, only area-specific auxiliary data are available for the population elements while in the second type element-specific auxiliary data are available for the population elements.

We have emphasized the importance of obtaining accurate measures of uncertainty associated with the model-based estimators. To this end, an approximately unbiased estimator of MSE of the EBLUP estimator is given as well as two methods of approximating the true posterior variance, irrespective of the form of the prior distribution on the model parameters. The latter approximations may be used as measures of uncertainty associated with the EB estimator. In the HB approach, a prior distribution on the model parameters is specified and the resulting posterior variance is used as a measure of uncertainty associated with the HB estimator (posterior mean). We have also mentioned several applications of the model-based methods.

We have also considered special problems that may be encountered in implementing model-based methods for small area estimation; in particular, model diagnostics for small area models, constrained estimation, "local" shrinkage, spatial modelling and borrowing strength across both small areas and time. We anticipate quite a bit of future research on these topics.

Caution should be exercised in using or recom-

mending indirect estimators since they are based on implicit or explicit models that connect the small areas, unlike the direct estimators. As noted by Schaible (1992): "Indirect estimators should be considered when better alternatives are not available, but only with appropriate caution and in conjunction with substantial research and evaluation efforts. Both producers and users must not forget that, even after such efforts, indirect estimates may not be adequate for the intended purpose." (Also see Kalton, 1987.)

Finally, we should emphasize the need for developing an overall program that covers issues relating to sample design and data evolvement, organization and dissemination, in addition to those pertaining to methods of estimation for small areas.

APPENDIX

Variances and Covariance of $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$

Let $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ be the estimators of σ_v^2 and σ^2 obtained from the method of fitting constants. Then

$$V(\hat{\sigma}^2) = 2\nu_1^{-1}\sigma^4$$

$$V(\hat{\sigma}_v^2) \doteq 2\eta_*^{-2}\Big[\nu_1^{-1}(n-p-\nu_1)(n-p)\sigma^4 + \eta_{**}\sigma_v^4 + 2\eta_*\sigma^2\sigma_v^2\Big]$$

with

$$\eta_{**} = \sum w_{i\cdot}^2\left(1 - w_{i\cdot}\bar{\mathbf{x}}_{iw}^T\mathbf{A}_1^{-1}\bar{\mathbf{x}}_{iw}\right) + \operatorname{tr}\left(\mathbf{A}_1^{-1}\sum w_{i\cdot}^2\bar{\mathbf{x}}_{iw}\bar{\mathbf{x}}_{iw}^T\right)^2$$

and

$$\operatorname{cov}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \doteq -2\eta_*^{-1}\nu_1^{-1}(n-p-\nu_1)\sigma^4.$$

(See Stukel, 1991.)

ACKNOWLEDGMENTS

The authors would like to thank the former editor, J. V. Zidek, and a referee for several constructive suggestions. M. Ghosh was supported by NSF Grants DMS-89-01334 and SES-92-01210. J. N. K. Rao's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

ANDERSON, T. W. and HSIAO, C. (1981). Formulation and estimation of dynamic models using panel data. *J. Econometrics* 18 67–82.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* 83 28–36.

BECKMAN, R. J., NACHTSHEIM, C. J. and COOK, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics* 29 413–426.

BOGUE, D. J. (1950). A technique for making extensive postcensal estimates. *J. Amer. Statist. Assoc.* 45 149–163.

BOGUE, D. J. and DUNCAN, B. D. (1959). A composite method of estimating post censal population of small areas by age, sex and colour. Vital Statistics–Special Report 47, No. 6, National Office of Vital Statistics, Washington, DC.

BRACKSTONE, G. J. (1987). Small area data: policy issues and technical challenges. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh eds.) 3–20. Wiley, New York.

CALVIN, J. A. and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *J. Amer. Statist. Assoc.* 86 36–48.

CHAMBERS, R. L. and FEENEY, G. A. (1977). Log linear models for small area estimation. Unpublished paper, Australian Bureau of Statistics.

CHAUDHURI, A. (1992). Small domain statistics: a review. Technical report ASC/92/2, Indian Statistical Institute, Calcutta.

CHOUDHRY, G. H. and RAO, J. N. K. (1989). Small area estimation using models that combine time series and cross-sectional data. In *Symposium 89—Analysis of Data in Time—Proceedings* (A. C. Singh and P. Whitridge, eds.) 67–74. Statistics Canada, Ottawa.

CHRISTENSEN, R., PEARSON, L. M. and JOHNSON, W. (1992). Case deletion diagnostics for mixed models. *Technometrics* 34 38–45.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *J. Amer. Statist. Assoc.* 84 1033–1044.

CRESSIE, N. (1990a). Small area prediction of undercount using the general linear model. In *Symposium 90—Measurement and Improvement of Data Quality—Proceedings* 93–105. Statistics Canada, Ottawa.

CRESSIE, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology* 18 75–94.

CRONKITE, F. R. (1987). Use of regression techniques for developing state and area employment and unemployment estimates. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 160–174. Wiley, New York.

DATTA, G. S. and GHOSH, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Ann. Statist.* 19 1748–1770.

DATTA, G. S., GHOSH, M., HUANG, E. T., ISAKI, C. T., SCHULTZ, L. K. and TSAY, J. H. (1992). Hierarchical and empirical Bayes methods for adjustment of census undercount: the 1980 Missouri dress rehearsal data. *Survey Methodology* 18 95–108.

DEMPSTER, A. P. and RYAN, L. M. (1985). Weighted normal plots. *J. Amer. Statist. Assoc.* 80 845–850.

DESOUZA, C. M. (1992). An approximate bivariate Bayesian method for analysing small frequencies. *Biometrics* 48 1113–1130.

DREW, D., SINGH, M. P. and CHOUDHRY, G. H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology* 8 17–47.

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimates—Part II: the empirical Bayes case. *J. Amer. Statist. Assoc.* 67 130–139.

ERICKSEN, E. P. (1974). A regression method for estimating populations of local areas. *J. Amer. Statist. Assoc.* 69 867–875.

ERICKSEN, E. P. and KADANE, J. B. (1985). Estimating the population in a census year (with discussion). *J. Amer. Statist. Assoc.* 80 98–131.

ERICKSEN, E. P. and KADANE, J. B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C .E. Särndal and M. P. Singh, eds.) 23–45. Wiley, New York.

ERICKSEN, E. P. and KADANE, J. B. (1992). Comment on "Should we have adjusted the U.S. Census of 1980?," by D. A. Freedman and W. C. Navidi. *Survey Methodology* 18 52–58.

ERICKSEN, E. P., KADANE, J. B. and TUKEY, J. W. (1989). Adjusting the 1981 census of population and housing. *J. Amer. Statist. Assoc.* 84 927–944.

ERICSON, W. A. (1969). A note on the posterior mean. *J. Roy. Statist. Soc. Ser. B* 31 332–334.

FARRELL, P. J., MACGIBBON, B. and TOMBERLIN, T.J. (1992). An evaluation of bootstrap techniques for correcting empirical Bayes interval estimates. Unpublished manuscript, Dept. Statistics and Actuarial Science, Univ. Waterloo.

FAY, R. E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 91–102. Wiley, New York.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* 74 269–277.

FREEDMAN, D. A. and NAVIDI, W. C. (1986). Regression models for adjusting the 1980 Census (with discussion). *Statist. Sci.* 1 1–39.

FREEDMAN, D. A. and NAVIDI, W. C. (1992). Should we have adjusted the U.S. Census of 1980? (with discussion). *Survey Methodology* 18 3–74.

FULLER, W. A. and BATTESE, G. E. (1973). Transformations for estimation of linear models with nested error structure. *J. Amer. Statist. Assoc.* 68 626–632.

FULLER, W. A. and HARTER, R. M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 103–123. Wiley, New York.

GELFAND, A. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 398–409.

GHANGURDE, D. D. and SINGH, M. P. (1977). Synthetic estimators in periodic households surveys. *Survey Methodology* 3 152–181.

GHOSH, M. (1992a). Hierarchical and empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh and P. K. Pathak, eds.) 151–177. IMS, Hayward, CA.

GHOSH, M. (1992b). Constrained Bayes estimation with applications. *J. Amer. Statist. Assoc.* 87 533–540.

GHOSH, M., DATTA, G. S. and FAY, R. E. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. In *Proceeding of the Bureau of the Census Annual Research Conference* 63–79. Bureau of the Census, Washington, DC.

GHOSH, M. and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* 82 1153–1162.

GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimators. In *Proceedings of the Social Statistics Section* 33–36. Amer. Statist. Assoc., Washington, DC.

HANSEN, M., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*, 1. Wiley, New York.

HARVILLE, D. A. (1991). Comment on, "That BLUP is a good thing: The estimation of random effects," by G. K. Robinson. *Statist. Sci.* 6 35–39.

HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* 21 309–310.

HOLT, D., SMITH, T. M. F. and TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *J. Amer. Statist. Assoc.* 74 405–410.

HULTING, F. L. and HARVILLE, D. A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small area estimation: computational aspects, frequentist properties, and relationships. *J. Amer. Statist. Assoc.* 86 557–568.

ISAKI, C. T., SCHULTZ, L. K., SMITH, P. J. and DIFFENDAL, D. J. (1987). Small area estimation research for census undercount—progress report. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C.E. Särndal and M.P. Singh, eds.) 219–238. Wiley, New York.

JUDGE, G. G. (1985). *The Theory and Practice of Econometrics*. 2nd ed. Wiley, New York.

KACKAR, R. N. and HARVILLE, D. A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *J. Amer. Statist. Assoc.* 79 853–862.

KALTON, G. (1987). Discussion. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 264-266. Wiley, New York.

KALTON, G., KORDOS, J. and PLATEK, R. (1993). *Small Area Statistics and Survey Designs Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion*. Central Statistical Office, Warsaw.

KASS, R. E. and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* 84 717–726.

KLEFFE, J. and RAO, J. N. K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *J. Multivariate Anal.* 43 1–15.

LAHIRI, P. (1990). "Adjusted" Bayes and empirical Bayes estimation in finite population sampling. *Sankhyā Ser. B* 52 50–66.

LAHIRI, P. and RAO, J. N. K. (1992). Robust estimation of mean square error of small area estimators. Unpublished manuscript.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* 82 739–750.

LANGE, N. and RYAN, L. (1989). Assessing normality in random effects models. *Ann. Statist.* 17 624–642.

LOUIS T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* 79 393–398.

MACGIBBON, B. and TOMBERLIN, T.J. (1989). Small area estimation of proportions via empirical Bayes techniques. *Survey Methodology* 15 237–252.

MALEC, D., SEDRANSK, J. and TOMPKINS, L. (1991). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. Unpublished manuscript.

MARKER, D. A. (1983). Organization of small area estimators. *Proceedings of Survey Research Methods Section* 409–414. Amer. Statist. Assoc., Washington, DC.

MARSHALL, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *J. Roy. Statist. Soc. Ser. C* 40 283–294.

MCCULLAGH, P. and ZIDEK, J. (1987). Regression methods and performance criteria for small area population estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 62–74. Wiley, New York.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussions). *J. Amer. Statist. Assoc.* 78 47–65.

MORRISON, P. (1971). Demographic information for cities: a manual for estimating and projecting local population characteristics. RAND report R-618-HUD.

NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Synthetic State Estimates of Disability*. P.H.S. Publication 1759. U.S. Government Printing Office, Washington, DC.

NATIONAL INSTITUTE ON DRUG ABUSE (1979). *Synthetic Estimates*

for Small Areas (research monograph 24). U.S. Government Printing Office, Washington, DC.

NATIONAL RESEARCH COUNCIL (1980). *Panel on Small-Area Estimates of Population and Income. Estimating Population and Income of Small Areas.* National Academy Press, Washington, DC.

NICHOL, S. (1977). A regression approach to small area estimation. Unpublished manuscript, Australian Bureau of Statistics, Canberra, Australia.

NTIS (1963). Indirect estimators in federal programs. Statistical policy working paper 21, prepared by the Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC.

PFEFFERMANN, D. and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* 16 217–237.

PFEFFERMANN, D. and BARNARD, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics* 9 73–84.

PLATEK, R. and SINGH, M. P. (1986). *Small Area Statistics: Contributed Papers.* Laboratory for Research in Statistics and Probability, Carleton Univ.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E. and SINGH, M. P. (1987). *Small Area Statistics.* Wiley, New York.

PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of mean squared errors of small-area estimators. *J. Amer. Statist. Assoc.* 85 163–171.

PURCELL, N. J. and LINACRE, S. (1976). Techniques for the estimation of small area characteristics. Unpublished manuscript.

PURCELL, N. J. and KISH, L. (1979). Estimation for small domain. *Biometrics* 35 365–384.

PURCELL, N. J. and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Internat. Statist. Rev.* 48 3–18.

RAO, C. R. and SHINOZAKI, N. (1978). Precision of individual estimates in simultaneous estimation of parameters. *Biometrika* 65 23–30.

RAO, J. N. K. (1986). Synthetic estimators, SPREE and best model based predictors. In *Proceedings of the Conference on Survey Research Methods in Agriculture* 1–16. U.S. Dept. Agriculture, Washington, DC.

RAO, J. N. K. and YU, M. (1992). Small area estimation by combining time series and cross-sectional data. In *Proceedings of the Survey Research Methods Section* 1–19. Amer. Statist. Assoc., Alexandria, VA.

ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statist. Sci.* 6 15–51.

SÄRNDAL, C. E. and HIDIROGLOU, M. A. (1989). Small domain estimation: a conditional analysis. *J. Amer. Statist. Assoc.* 84 266–275.

SCHAIBLE, W. L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the Survey Research Methods Section* 741–746. Amer. Statist. Assoc., Washington, DC.

SCHAIBLE, W. L. (1992). Use of small area statistics in U.S. Federal Programs. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) 1 95–114. Central Statistical Office, Warsaw.

SINGH, A. C. and MANTEL, H. J. (1991). State space composite estimation for small areas. In *Symposium 91—Spatial Issues in Statistics—Proceedings* 17–25. Statistics Canada, Ottawa.

SINGH, M. P., GAMBINO, J. and MANTEL, H. (1992). Issues and options in the provision of small area data. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) 1 37–75. Central Statistical Office, Warsaw.

SMITH, S. K. and LEWIS, B. B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography* 17 323–340.

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.

STARSINIC, D. E. (1974). Development of population estimates for revenue sharing areas. Census Tract Papers, Ser. GE40, No. 10 U.S. Government Printing Office, Washington, DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*, Catalogue 91-528E. Statistics Canada, Ottawa.

STEFFEY, D. and KASS, R. E. (1991). Comment on "That BLUP is a good thing: The estimation of random effects," by G. K. Robinson. *Statist. Sci.* 6 45–47.

STUKEL, D. (1991). *Small Area Estimation Under One and Two-Fold Nested Error Regression Model.* Ph.D. Thesis, Carleton Univ.

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

ZIDEK, J. V. (1982). A review of methods for estimating the populations of local areas. Technical Report 82–4, Univ. British Columbia, Vancouver.

Comment

Noel Cressie and Mark S. Kaiser

Malay Ghosh and Jon Rao have presented us with a well written exposition of the topic of small area estimation. The past literature has been decidedly influenced by linear modeling, and we see that clearly in their paper. There has also been a tendency to judge the performance of the estimation methods by concentrating on a single, arbitrary small area. In our comment, we shall discuss what opportunities there might be to expand the class of statistical models for small area data and to consider multivariate aspects of small area estimation.

Noel Cressie is Professor of Statistics and Distinguished Professor in Liberal Arts and Sciences and Mark S. Kaiser is Assistant Professor, Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa 50011-1201.

MODELING APPROPRIATE SOURCES OF VARIABILITY

It would appear from the authors' account that the full flexibility of hierarchical modeling has not been applied to small area estimation. Two models incorporating random effects, given as equations (4.4) and (4.5), are presented in their paper. Model (4.4) is applied when both direct estimators and auxiliary data are available at the area level while model (4.5) is partitioned into sampled and unsampled units within a small area when both response and auxiliary data are available for sampling units. In either case, estimates of the area means or totals are developed. Rather than focusing on this distinction, we would like to point out the similarity of these models in the way that additional response variability is due to the random nature of model components. Using y as a vector of response data, both of the models may be considered hierarchical models of the general form,

$$(1)\quad [y \mid \mu, \Sigma] = N(\mu, \Sigma) \quad \text{and} \quad [\mu \mid \beta, \Gamma] = N(X\beta, \Gamma),$$

where $[y \mid \theta]$ denotes the probability distribution of y given the parameter θ; and both Σ and Γ are positive-definite matrices. Then the marginal distribution of y is immediately (e.g., Lindley and Smith, 1972, Lemma 1)

$$(2)\qquad N(X\beta, \Sigma + \Gamma),$$

which also results from writing the models in mixed linear form (as Ghosh and Rao have chosen to do). The covariance matrix of the marginal density indicates that these types of models incorporate *sampling variability* into the distribution of y through the use of hierarchical structure. (In engineering, this approach is called state-space modeling.) What might be considered the *systematic* model component, namely $E(y) = X\beta$, is generated in the same way across all areas in model (4.4) or across all sampling units within all areas in model (4.5).

The hierarchical model described by (1) is different from the model,

$$(3)\quad [y \mid \beta, \Sigma] = N(X\beta, \Sigma) \quad \text{and} \quad [\beta \mid B, \Gamma] = N(B, \Gamma),$$

for which the marginal density of y becomes

$$(4)\qquad N(XB, \Sigma + X\Gamma X^T).$$

Under the hierarchical model (3), variability in the marginal distribution of y is affected by the values of the explanatory variables observed.

A third possibility suggests itself in the situation that unit specific observations are available in each of several small areas. In this case, one might apply model (3) to each area using y_i, β_i, X_i and Σ_i to denote the dependence on area identification. The $\{\beta_i\}$ could be taken as independent and identically distributed random variables with common distribution across areas; for example, $[\beta_i \mid B, \Gamma] = N(B, \Gamma)$. Then, with assumed independence (conditional on $\{\beta_i, \Sigma_i\}$) of the $\{y_i\}$, the joint marginal of all observations is available as a product of the marginals for the m areas. More complex models allowing lack of independence for either the y_i or β_i are conceivable; and, in fact, model (3) is an example of one such. Under models of this general type, variability among observations comes not only from direct sampling variability but also from variability in the $\{\beta_i\}$ that describe the systematic relation between y and X. That such variability often exists seems a reasonable supposition. In the introductory example discussed by Ghosh and Rao, of estimating per capita income (PCI) for local administrative areas (Fay and Herriot, 1979), a regression of estimated PCI on county tax returns and housing data is assumed. The systematic relation described by such a regression may well be different for counties in different portions of a state or region, as may be the range and values of the explanatory variables used. As another example, the small areas where census undercounts are estimated can each be stratified by race. A separate regression for each race (Cressie, 1989) results in differences in regression coefficients. Finally, estimation of the distribution of regression coefficients may provide valuable information to demographers and social scientists, such as in the problem of census undercount.

There is a difference in the modeling approach represented by (1) on the one hand, and (3) and its extensions on the other, that centers on the sources of variation in the observed responses. From a Bayesian viewpoint, this difference involves the order in which prior distributions are placed on model components. The order in which priors are assigned is pertinent, particularly in light of the fact that the data contain less information about parameters as those parameters move up in the hierarchy (Goel and DeGroot, 1981). Thus, if we have interest in the posterior distribution of β, we are well served by positioning β low in the hierarchy which leads to model (3) and its extensions rather than to model (1). Under model (1), we do not question the strength of the linear relation between y and X but are uncertain about the realization that may be observed in any particular small area. Then, a completely specified, but often uninformative, prior is placed on β as much to allow computation of a posterior distribution of μ as from genuine interest in modeling either prior or posterior distributions of β. Under

model (3) and its extensions, an important source of uncertainty stems from lack of knowledge about β or $\{\beta_i\}$. Ghosh and Rao have not discussed the latter approach in small area estimation and it would be interesting and useful to see what differences might result from its application. The extension of model (3) to area-specific regression equations, in particular, offers an interesting alternative to the standard approach in that it raises the possibility of predicting the area-specific regression parameters $\{\beta_i\}$.

NONLINEAR MODELS

In an effort to increase the flexibility of small area models, it is natural to consider ways to extend the modeling concepts to nonlinear situations. One approach to nonlinear modeling that encompasses many situations is that of generalized linear models (GLMs). Ghosh and Rao mention binary and Poisson responses in Section 7 of their paper which fall into this framework. Our earlier discussion of appropriate sources of variability carries over to the GLM, and we give several models analogous to the normal models already presented. While the notation of GLMs offers flexibility in allowing nonlinear response functions, there is a concomitant reduction in flexibility for modeling lack of independence among responses. Specifically, small area responses y_i are taken to be univariate random variables, and conditional independence of these variables (conditional on parameters) is assumed throughout. Assume that y_i is distributed according to an exponential family with density (or mass) function,

$$f(y_i \mid \theta_i, \phi_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi_i) + c(y_i, \phi_i)\},$$

so that $E(y_i) = b'(\theta_i) \equiv \mu_i$ and $\operatorname{var}(y_i) = a(\phi_i)b''(\theta_i) \equiv a(\phi_i)V(\mu_i)$. A GLM is completed by taking a known function of μ_i to be linear in a set of covariates; that is, $g(\mu_i) = x_i^T\beta \equiv \eta_i$ with $x_i = (x_{i1}, \ldots, x_{ip})^T$. One hierarchical extension of this model is to let the natural parameter θ_i be distributed according to some probability density (or mass) function $h(\theta_i \mid \lambda)$. The marginal density (or mass) function of y_i then becomes

$$(5) \qquad p(y_i \mid \lambda, \phi_i) = \int f(y_i \mid \theta_i, \phi_i)h(\theta_i \mid \lambda)d\theta_i.$$

This is the approach taken by Albert (1988) and Albert and Pepple (1989) to develop hierarchical overdispersion models. These authors take $h(\theta_i \mid \lambda)$ from a conjugate exponential family for f and then set up the GLM by linking the expected value of μ_i with a linear model as $g(E(\mu_i)) = x_i^T\beta$. This approach moves the linear model away from y to a position further up in the hierarchy and is analogous to the approach of model (1) leading exactly to that model in the case that y is normal with mean $\theta = \mu$, and g the identity mapping.

A different approach, analogous to that used in model (3), is to start with a fully specified GLM for the responses and allow β to be random. In this case, we must assign a multivariate distribution for β. For example, we might take $[\beta \mid B, \Gamma] = N(B, \Gamma)$. The marginal distribution of y_i is then

$$(6) \quad p(y_i \mid B, \Gamma, \phi_i) = \int f(y_i \mid \beta, \phi_i)h(\beta \mid B, \Gamma)d\beta.$$

In (6), we use the exponential form for y_i and the systematic specification $g(\mu_i) = x_i^T\beta$, so that $\theta_i = b'^{-1}[g^{-1}(x_i^T\beta)]$ and

$$(7) \quad \begin{aligned} &f(y_i \mid \beta, \phi_i) \\ &\quad = \exp\{[y_i b'^{-1}(g^{-1}(x_i^T\beta)) \\ &\qquad - b[b'^{-1}(g^{-1}(x_i^T\beta))]]/a(\phi_i) + c(y_i, \phi_i)\}. \end{aligned}$$

Things simplify substantially by taking g as the canonical link function $g(\cdot) = b'^{-1}(\cdot)$, giving

$$(8) \quad \begin{aligned} &f(y_i \mid \beta, \phi_i) \\ &\quad = \exp\{[y_i x_i^T\beta - b(x_i^T\beta)]/a(\phi_i) + c(y_i, \phi_i)\}. \end{aligned}$$

Using expression (8), it would be possible, at least in theory, to complete the integrations in (6). In practice, the necessary integrations might be best approached through importance sampling or, in a Bayesian analysis, the joint posterior distribution might be calculated directly through Monte Carlo resampling schemes (e.g., Smith and Roberts, 1993).

As for the analogous normal model (3), the ideas culminating in equation (8) may be extended directly to the situation of different regressions among areas. Unlike the normal situation, however, it is difficult to conceptualize the way lack of independence among responses in different small areas could be handled.

In all of these hierarchical models, ways to deal with dispersion parameters and covariance matrices become a major statistical issue. It is always possible, in theory, to find maximum likelihood estimators. Can small sample properties of maximum likelihood estimators of dispersion and covariance parameters be improved, perhaps using some appropriate analogues to REML estimation? Bayesian modeling of these (nuisance) parameters is another possibility that is becoming feasible with the recent developments in Gibbs sampling and Monte Carlo resampling schemes.

MULTIVARIATE ASPECTS

Although the problem of small area estimation is inherently multivariate, there has been a tendency to look at the performance of estimation procedures area-by-area. For example, Ghosh and Rao give the mean-squared error formula (5.5) for the ith small area. What is actually needed is the $m \times m$ mean-squared error matrix

$$(9) \qquad (\mathrm{mse}(i,j)) = E\{(\hat{\theta}^H - \theta)(\hat{\theta}^H - \theta)^T\},$$

whose diagonal elements are given by (5.5) but whose off-diagonal elements also have an important role to play.

Suppose that two small areas i and i' are combined into a new area that we denote $i \cup i'$. Now, assuming a linear model, $\theta_{i \cup i'} = w\theta_i + w'\theta_{i'}$ and $\hat{\theta}^H_{i \cup i'} = w\hat{\theta}^H_i + w'\hat{\theta}^H_{i'}$. Hence,

$$\mathrm{mse}(i \cup i', i \cup i') = w^2\,\mathrm{mse}(i,i) + (w')^2\,\mathrm{mse}(i',i') + 2ww'\,\mathrm{mse}(i,i'),$$

which involves both diagonal and off-diagonal elements of (9). Cressie (1992) develops an approximation to (9), analogous to the univariate approximation (5.5).

As another example, a multivariate version of the Laird and Louis (1987) bootstrap, described by the authors in Section 5.2, is straightforward to derive. Let $\theta \equiv (\theta_1, \ldots, \theta_m)$ denote the parameters of the m small areas. A large number, B, of independent bootstrap samples $\{\theta^*(b) : b = 1, \ldots, B\}$ are drawn from the estimated marginal distribution $N(X\hat{\beta}, \hat{\Sigma} + \hat{\Gamma})$; see relation (2). Estimates $\beta^*(b), \Sigma^*(b)$ and $\Gamma^*(b)$ are computed from the bootstrap data $\theta^*(b)$ for each b. Then the EB bootstrap estimator and the appropriate estimated posterior variance matrix are, respectively,

$$\theta^{*EB}(\cdot) = (1/B)\sum_{b=1}^{B} E(\theta \mid \theta^*(b), \beta^*(b), \Sigma^*(b), \Gamma^*(b)) = (1/B)\sum_{b=1}^{B} \theta^{*EB}(b),$$

$$V^* = (1/B)\sum_{b=1}^{B} \mathrm{var}(\theta \mid \theta^*(b), \beta^*(b), \Sigma^*(b), \Gamma^*(b)) + (1/(B-1))\sum_{b=1}^{B} (\theta^{*EB}(b) - \theta^{*EB}(\cdot)) \cdot (\theta^{*EB}(b) - \theta^{*EB}(\cdot))^T.$$

Given the geographic nature of most small area estimation problems, the question of how to aggregate is always waiting to be asked; hence, the multivariate aspects are important. The harder question of how to disagregate has been at the core of much of the debate about the adjustment of census counts. Cressie (1988) shows that adjustment based on small area estimation of both the synthetic and empirical Bayes type offers smaller risk than no adjustment even under disaggregation of the small areas. Crucial to his argument is the appropriateness of the small area model at the disaggregated level. Tukey (1983) and Wolter and Causey (1991) reach similar conclusions to Cressie; however, both articles make an assumption that when disaggregating synthetically the *true* adjustment factor is *known* at the level below which disaggregation occurs. There is no certainty that adjustment will improve counts at *all* disaggregated levels; Freedman and Navidi (1992) give a simple example to demonstrate that some adjusted counts can be worse than unadjusted counts.

CONSTRAINED ESTIMATION

In a sense, constrained estimation takes a multivariate point of view in that interest is focussed on how well the *ensemble* of the m small area estimators matches the ensemble of the m estimands. However, there is an opportunity to make the problem more explicitly multivariate.

First, we would like to fill in some of the history of constrained estimation. Tukey (1974, page 143) was aware that the ensemble of estimates gives poor information about the ensemble properties of parameters (e.g., one such property might be the population-weighted proportion of small areas whose lip-cancer rate is above .05 per thousand population years at risk). Louis (1984) addressed the problem in a normal homoscedastic model by advocating that optimal (i.e., Bayes) shrinkage estimates be modified so that their ensemble variance matches the posterior expectation of the parameters' ensemble variance. Cressie (1986, 1989) coined the term "constrained Bayes estimation" and generalized Louis' result to heteroscedastic normal models (for census undercount).

Spjøtvoll and Thomsen (1987) completely ignored the multivariate aspects of the problem by considering each area one-at-a-time. Let θ_i and $\bar{\theta}_i$ denote the parameter and an estimator, respectively, for the ith area. Assume that both parameter and estimator are random, with first two moments finite, and that $E(\bar{\theta}_i \mid \theta_i) = \theta_i$. They propose to estimate θ_i by

$$(10) \qquad \hat{\theta}_i = a_i\bar{\theta}_i + b_i,$$

where a_i and b_i are solved by specifying that $E(\hat{\theta}_i) = E(\theta_i) \equiv \nu$ and $\mathrm{var}(\hat{\theta}_i) = \mathrm{var}(\theta_i) \equiv \sigma^2$. In the discussion to Spjøtvoll and Thomsen's paper, it is pointed

out that the solution yields the constrained empirical Bayes estimates obtained by Cressie (1986), although no Bayes optimality criterion is invoked by the authors.

The multivariate version of (10) is

$$\hat{\theta} = A\tilde{\theta} + b, \tag{11}$$

where A is an $m \times m$ matrix and b is an $m \times 1$ vector. Upon specifying that $E(\hat{\theta}) = E(\theta)$ and $\text{var}(\hat{\theta}) = \text{var}(\theta)$, Cressie (1990b, 1992) obtains a multivariate constrained estimator. In the notation of (1), $\theta = \mu$, $E(\theta) = X\beta$, $\tilde{\theta} = y$, $E(y \mid \theta) = \theta$, $\text{var}(y \mid \theta) = \Sigma$, and $\text{var}(y) = \Sigma + \Gamma$. Then the multivariate constrained estimator for model (1), analogous to Spjøtvoll and Thomsen's, is given by (11), where

$$A = \Gamma^{1/2}(\Sigma + \Gamma)^{-1/2} \tag{12}$$

and

$$b = \{I - \Gamma^{1/2}(\Sigma + \Gamma)^{-1/2}\}X\beta. \tag{13}$$

Notice that $\hat{\theta}$ given by (11), (12) and (13) does not shrink y towards $X\beta$ as far as the Bayes estimator θ^* does (where $A = \Gamma(\Sigma + \Gamma)^{-1}$ and $b = (I - A)X\beta$).

In an elegant paper, Ghosh (1992) derives a multivariate constrained *Bayes* estimator for model (1):

$$\theta^{@} = \{a + (1-a)\underset{\sim}{1}\underset{\sim}{1}'/m\}\theta^*, \tag{14}$$

where

$$a = \left[\text{trace}\{(I - \underset{\sim}{1}\underset{\sim}{1}'/m)V\}\left(\sum_{i=1}^{m}(\theta_i^* - \bar{\theta}^*)^2\right)^{-1} + 1\right]^{1/2},$$

$$\theta^* = E(\theta \mid y) = \{\Gamma(\Sigma + \Gamma)^{-1}\}y + (I - \Gamma(\Sigma + \Gamma)^{-1}\}X\beta,$$

$$V = \text{var}(\theta \mid y) = \Gamma\{I - \Gamma(\Sigma + \Gamma)^{-1}\}\Gamma.$$

The vector $\theta^{@}$ has the property that it minimizes $E(\Sigma_{i=1}^{m}(\theta_i - t_i)^2 \mid y)$ with respect to t and subject to conditions that match first and second sample moments of t with those same moments of θ conditional on y. Cressie's proposal given by (11), (12) and (13) does not invoke any optimality conditions and so is likely to be less efficient than Ghosh's estimator (14).

Constrained Bayes estimation for more general models, such as GLMs, is presented by Ghosh (1992), although from an essentially univariate point of view. Our earlier comment, that we do not have flexible ways to model lack of independence in nonlinear, nonnormal models, is equally appropriate here.

Finally, we agree with the authors' comment about the importance of small area estimation in medical geography. A good source for recent research in this area is the May 1993 Supplement Issue of the journal *Medical Care* (Proceedings of the Fourth Biennial Regenstrief Conference, "Methods for Comparing Patterns of Care," October 27–29, 1991). We are working on incorporating spatial variation and dependence into statistical methods for these and other small area estimation problems.

ACKNOWLEDGMENT

Partial support came from the Office of Naval Research Grant N00014-93-1-0001, NSF Grant DMS-92-04521 and the National Security Agency Grant MDA904-92-H-3021.

Comment

D. Holt

The paper by Ghosh and Rao is a valuable summary of recent developments using empirical Bayes and hierarchical Bayes methods for making small area estimates. The need for methods which make provision for local variation while pooling information across areas is well established. The review is a thorough appraisal of the methods and their properties, and the numerical results reinforce earlier results which demonstrate that these methods are preferable to others such as synthetic estimation and sample size dependent estimation.

The value of these approaches is not simply in their ability to provide point estimates for each small area which, on average, have better precision. A very important additional factor is that a measure of precision (MSE) and an estimator of this can be

D. Holt is Professor, Department of Social Statistics, University of Southampton, Southampton S09 5NH, United Kingdom.

developed for each small area separately. This is extremely important since the precision of each small area estimate will depend upon a number of factors including the sample size and the distribution of the area specific covariate values as well as the method of estimation itself. Indeed, no one method of estimation will be necessarily uniformly superior for all small areas; and any choice of estimator will result in a loss of precision for some areas as well as gains for others.

This point leads to the issue of which measures of precision are appropriate and how to present numerical results. Ghosh and Rao present a single point estimate for each small area from a single sample. It is, in effect, a simulation of size one. There are advantages to this approach since we can make direct comparisons between the estimates and true small area means in each case. I will return to this point, but first let us consider the numerical results as presented.

The choice of measures, relative absolute error and squared error for each small area separately, are both natural. The first represents a measure analogous to coefficient of variation and the second represents MSE. However, it is dangerous to summarize these measures into a single average across all small areas without paying some attention to the distribution. In Table 3, for example, one notices that for all four estimators considered the point estimates are less than the true values for 13 of the 16 small areas. Also for each estimator the two measures are extremely variable across the small areas. To consider the sample dependent estimator, for example, Ghosh and Rao comment that in terms of average relative error it is similar to EBLUP and HB but in terms of average squared error it is inferior. However, one may derive from the table that 58% of the ASE for this estimator is derived from the last small area. For each of the estimators, the distribution of relative absolute error and squared error is informative and important.

When one considers the distribution of performance measures for each small area, then the reader cannot separate systematic performance from random error since the results represent a simulation of size one. Is it the case, for example, that in small area 4 the tiny deviations for the ratio synthetic and sample size dependent estimators and the much larger deviations for EBLUP and HB reflect a true difference in performance or is this random fluctuation? Would it not have been better to produce measures which were based upon a set of repeated simulations and which could have included an average bias, average relative absolute error and mean squared error for each small area separately? If this had been done then comparisons could have been made between estimators for each small area separately (e.g., comparison of average bias, MSE, etc.). The distribution of these comparative measures and their overall summary could then have been considered.

This rather simple comment raises a rather fundamental issue about the framework for measures of performance and how numerical simulations should be designed. The measures of precision (e.g., MSE) given in Section 5 of Ghosh and Rao's paper are essentially model based. To some extent assumptions of normality are required but the authors comment about the robustness of the methods. However the properties within the model framework are conditional on the values of the auxiliary variable (x_{ij}) and the sample size achieved in each small area (n_i). Within the predictive framework many analysts would prefer measures of precision which condition on the achieved sample in this way. Survey practitioners, on the other hand, and anyone considering the choice of estimation method in advance of the survey being analyzed will want to understand the properties of estimators across of range of circumstances. This creates a dilemma for the theoretician who wishes to demonstrate the comparative properties of alternative estimation methods, using simulations.

Should Ghosh and Rao:

(a) Fix the sample values of n_i, $\{x_{ij}\}$ and a single randomly generated value of the small area effect ν_i and carry out repeated simulations to obtain the properties of each estimator for each small area?
(b) Fix the sample values of n_i and simulate repeated sample selections from the population of each small area?
(c) Draw repeated random samples from the whole population without restriction?

The model based MSE given in Section 5 will be constant under (a) but not under (b) or (c). An analyst might be more interested in (a) but would want to be assured that the results did not depend on the particular choice of the sample configuration. Many survey practitioners would lean towards (b) or (c). Perhaps the practical solution is to draw several samples under (b) or (c); and then for each one selected, carry out repeated simulations under (a). By presenting the results from one simulation, Ghosh and Rao effectively avoid all of these issues.

Finally, to turn to a separate issue, the models described in Section 4 provide for local differences in the small area means by introducing a random effect, ν_i, for each small area. This is a random term which is the same for all units in the small area and essentially introduces a random effect for the

intercept of the linear model. This approach may be extended and the two model frameworks for equations (4.1) and (4.2) essentially integrated. Equation (4.5) may be generalized to allow all (or any) of the regression coefficients including the intercept to be random. Furthermore, small area level variables (z_i) may be used to explain some of the between small area variation:

$$y_i = x_i\beta_{1i} + e_i,$$
$$\beta_{1i} = z_i\gamma + \nu_i;$$

X_i is the $N_i \times (p+1)$ matrix of unit level covariates (including an intercept) and z_i is the $(p+1)\times q$ matrix of small area level variables. Here γ is the vector of length q of fixed coefficients and $\nu_i = (\nu_{i0}, \ldots, \nu_{ip})^T$ is a vector of length $p+1$ of random effects for the ith small area. In the general form the ν_i are independent between small areas but may have a joint distribution within each small area with $E(\nu_i) = 0$ and $V(\nu_i) = \Omega$:

$$\Omega = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0p} \\ \sigma_{10} & \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & & & \\ \sigma_{p0} & \sigma_{p1} & \cdots & \sigma_p^2 \end{bmatrix}$$

A special case is when the random effects are uncorrelated so that Ω is diagonal.

The use of area level variables, Z_i, to help explain the between area variation should help when the sample size in a small area is small. Also this more general model effectively integrates the use of unit level and area level covariates into a single model. Holt and Moura (1993) provide point estimates and expressions for MSE following the framework of Prasad and Rao (1990).

The use of extra random effects for the regression coefficients gives greater flexibility. If the unit level covariate is a set of dummy variables signifying group membership, for example, then this approach will allow a set of correlated and heteroscedastic random effects for the group means in each small area rather than a single random effect for all subjects.

The introduction of a random effect for the regression coefficient of a continuous covariate is likely to have more impact when the individual covariate values x_{ij} are variable within each small area. Judging by the values displayed in Table 2 where the values of x_{ij} vary greatly, it is possible that a more general model would provide even greater gains in precision for the empirical example which Ghosh and Rao consider.

Comment

Wesley L. Schaible and Robert J. Casady

Professors Ghosh and Rao have provided us with an excellent, comprehensive review of indirect estimation methods which have been suggested for the production of estimates for small areas and other domains. They make a timely contribution by reviewing and comparing a number of new methods which have recently appeared in the literature as well as updating previous work on some of the more established approaches. Demographic methods, synthetic and related estimators, empirical Bayes estimators, hierarchical Bayes estimators and empirical best linear unbiased prediction methods are thoroughly discussed; evidence that the Bayes and empirical prediction methods have advantages over the others is presented. Special problems in the application of small area estimation methods are also addressed. This is an extremely important issue and additional discussion would have been desirable. In our comments, we will expand on this subject by discussing some of the characteristics of indirect estimators and some specific practical problems associated with their use. In addition, we will attempt to state in general terms what we believe to be the fundamental problem associated with the application of small area estimation methodology.

Very generally speaking, applications of indirect estimation methods fall into one of three categories:

1. An indirect estimator is used to estimate a population parameter;
2. an indirect procedure is used to modify a direct estimator of a population parameter (e.g., a direct estimator that incorporates indirectly estimated post-stratification controls or seasonal

Wesley L. Schaible is Associate Commissioner and Robert J. Casady is Senior Mathematical Statistician, Office of Research and Evaluation, Bureau of Labor Statistics, 2 Massachusetts Avenue N.E., Washington, DC 20212.

adjustment procedures); and

3. an indirect estimator is used to estimate the variance of an estimator (e.g., a generalized variance function).

Essentially all of the small area estimation literature focuses on the first category of applications; the authors' review and our comments will do likewise.

The authors refer to the Federal Committee on Statistical Methodology report, "Indirect Estimators in Federal Programs." This report focuses on applications of indirect estimators and provides an interesting supplement to the paper under discussion. Some of the characteristics of indirect estimators and practical problems associated with their application, which are summarized in this report, are mentioned below:

- A domain and time specific model is implicitly assumed to be true when analyses among domains and over time are conducted. From a best linear unbiased prediction point of view, a domain and time specific model leads to a best linear unbiased direct estimator and also defines a family of indirect models which allow strength to be borrowed from other domains and/or time periods. The direct estimator is unbiased, not only under the domain and time specific model, but also under each of the corresponding indirect models. However, the best linear unbiased indirect estimators associated with the indirect models are not unbiased under the original domain and time specific model. This indirect estimator bias under the more plausible domain and time specific model adds to the uneasiness associated with the use of indirect estimators. It is the primary reason that indirect estimators are generally considered only when resources prohibit the use of direct estimators of adequate reliability.
- The variance of an indirect estimator will be smaller than that of the corresponding direct estimator since the indirect estimator not only incorporates observations of the variable of interest from the domain and time of concern but also from other domains and/or time periods.
- If the stochastic model underlying an indirect estimator is a satisfactory representation of reality, then the mean square error of the indirect estimator will likely be smaller than that of the corresponding direct estimator. However, many indirect estimators require strong model assumptions that may not be satisfied in most applications. If this is the case, then the mean square error of the indirect estimator may in fact be larger than the variance of the direct estimator. Although estimation of variances and (more importantly) mean square errors of indirect estimators has received attention, the estimation of a meaningful measure of error for a single small area remains a problem.
- Usually the task at hand is to produce estimates for a number of small areas simultaneously. There is considerable empirical evidence suggesting that the size of an error of an indirect estimator depends on the relationship of the area population value and population values of the other areas from which strength is borrowed. For example, the error in an indirect estimate for a small area with a very large population value is likely to be relatively large and negative, so that the estimate is closer to the population values of small areas that are not so large. This characteristic is not displayed to the same extent by all indirect estimators, and, as discussed by the authors, constrained estimators have been recently suggested to help address this problem.

The authors discuss the extremely important problem of model evaluation and suggest model diagnostics to help in the search for models that fit the data well. Until recently, model diagnostics have not played a major role in the evaluation of indirect estimators. Even though this approach is not free from dangers such as overfitting, especially when data sets are small, practitioners should make more use of these tools in estimator evaluation. Most government survey systems are designed to collect data and produce estimates periodically, yet the potential for continuing, routine estimator evaluations has not been fully explored. Problems of overfitting and small data sets associated with model diagnostics can be at least partially overcome by continuing evaluations.

We now turn to what we believe to be the general fundamental problem associated with the application of indirect estimation methods. A truly plausible model would depend on domain and time specific parameters, but indirect estimators are associated with models that contain one or more parameters that do not vary either over domains, time or both. In addition, in most practical applications, the statistician is pragmatically forced to settle for a stochastic model determined by the ancillary variables which are available. Models based on such expediency instill little confidence in either the producers or consumers of the estimates. Consequently, everyone concerned is usually convinced that the estimation process produces biased estimates; and, invariably, an empirical study is mounted to evaluate the average mean squared error (or some other ap-

propriate loss function) across the range of small areas. Such studies depend on "target values" for the parameter of interest for each small area, and generally accepted values of these target values are rarely, if ever, available (if they were, then there would be no need for indirect estimates). Thus, evaluation studies tend to produce conflicting and ambiguous results and leave all concerned less than completely satisfied. A good case in point are the many problems associated with use of a synthetic estimator to adjuste for state population undercounts in the 1990 census.

Having emphasized some of the problems associated with applications of indirect estimators, we should also mention the obvious fact that these estimation methods provide practitioners with many useful tools. Challenging research issues concerning the estimation of meaningful measures of error remain; without such measures, we must be cautious regarding inferences and actions based on these estimators. Nevertheless, in many applications, these methods provide us with an attractive alternative to the use of high variance direct estimates or, in some cases, no estimates at all.

Comment

Avinash C. Singh

The review paper of Ghosh and Rao fills a very important gap by giving a comprehensive and coherent picture of various developments in small area estimation over the last twenty years. This area is fascinating for at least three reasons: (1) there is a great demand for small area statistics by both government and private sectors for purposes of planning and policy analysis; (2) the small area problem provides a fertile ground for theoretical and applied research; and (3) the problem has attracted the attention of both Bayesians and frequentists because both approaches arise naturally and often seem to give similar results.

The main theme of my discussion is to compare and contrast the Bayesian and frequentist solutions to the problem of small area estimation. Why is it that for this problem the two approaches to statistical inference seem to converge in many practical examples including the one considered by Ghosh and Rao; that is, they provide similar results for both point estimates and the corresponding measures of uncertainty? Can we make some general statements about the similarity between the two approaches for small area estimation? How do their frequentist properties compare? Questions about the frequentist properties of some empirical Bayes methods are also raised by Ghosh and Rao in Section 5.2. Although the task of making exact comparisons is a difficult one, it is possible to make asymptotic comparisons for large m—the number of small areas. This will be the focus of my discussion.

Avinash C. Singh is Senior Methodologist, Methods Development and Analysis Section, Social Survey Methods Division, Statistics Canada, Ottawa K1A 0T6. He is also Adjunct Research Professor, Department of Mathematics and Statistics, Carleton University, Ottawa K1S 5B6.

1. MODEL REFORMULATION

As discussed in the review paper of Robinson (1991), understanding of procedures for estimating fixed and random effects helps to bridge the apparent gulf between the Bayesian and frequentist schools of thought. The present discussion will also strengthen this point. First, it will be convenient for our purposes to reformulate the model with fixed and random effects for small area estimation. Now, the general mixed linear model is given by

$$y = X\beta + Z\nu + \epsilon \tag{1}$$

where y is the n-vector of element-level data; X and Z are known matrices of orders $n \times p$ and $n \times m$, respectively, with rank $(X) = p$; β is a p-vector of fixed effects; ν is a m-vector of small area specific random effects and ϵ is a n-vector of random errors independent of ν such that $\nu \sim \text{WS}(0, G), \epsilon \sim \text{WS}(0, R)$. The abbreviation "WS" stands for "wide sense"; that is, the distribution is specified only up to the first two moments. The covariance matrices G and R depend on some parameters λ called variance components. For the reformulation of (1), we will regard the fixed effects β as random with mean 0 and covariance matrix $\sigma_\beta^2 I$ where $\sigma_\beta^2 \to \infty$. Thus, the limiting prior distribution of β is uniform (improper) which is commonly assumed in the Bayesian approach. The reformulation is useful for computational convenience as well as for making connections

between the Bayesian and frequentist approaches. Writing $\alpha = (\beta^T, \nu^T)^T$ and $F = (X, Z)$, we have the reformulated model,

$$y = F\alpha + \epsilon, \quad \alpha = \alpha^0 + \xi, \tag{2}$$

where $\alpha^0 = 0$, $\xi = (\beta^T, \nu^T)^T \sim \text{WS}(0, \Gamma), \Gamma) = \text{diag}(\sigma_\beta^2 I, G)$ and ξ is independent of ϵ.

The problem of interest is estimation (or prediction) of $L^T\alpha$ for some $(p+m)$-vector L. In the context of small areas, the vector L can be chosen appropriately to denote the superpopulation mean θ_i of each small area i. Note that if for each small area, population size is large and the sampling fraction is negligible, the estimation of finite population means is essentially equivalent to that of superpopulation means.

An important feature of the above reformulation [equation (2)] is that for known variance components λ, it provides a common model for both frequentist and Bayesian approaches. Not only does it provide a common starting point, both approaches yield identical estimates and the corresponding measures of uncertainty. Since the parameter of interest is inherently random in nature due to finiteness of the small area population, it is very appealing to have a unified formulation which gives identical results. However, for unknown λ, there is some divergence between the two approaches (see Section 3). First, we will consider the case of known λ.

2. CASE OF KNOWN VARIANCE COMPONENTS (λ KNOWN)

In this section, we show that when distributions are specified only in a wide sense, the Gauss-Markov theory (in the frequentist case) and the linear Bayes theory (in the Bayesian case) coincide. Under the frequentist approach for model (1), the objective is to find the best linear unbiased predictor (BLUP) of $\alpha = (\beta^T, \nu^T)^T$; that is, $\hat{\alpha} = a_0 + Ay$ is chosen to minimize

$$E\|a_0 + Ay - \alpha\|^2 \tag{3}$$

over all vectors a_0 and matrix A of appropriate dimensions. Here β is regarded as fixed and the expectation in (3) is with respect to y and ν. On the other hand, under the Bayesian approach, the objective is to find the (unbiased) linear Bayes estimate (LBE) of α as the prior information is specified in a wide sense only. The fixed effect β is assumed to have a uniform, improper prior distribution. Thus, the LBE $\tilde{\alpha} = A\alpha^0 + B(y - F\alpha^0)$ is obtained by minimizing

$$E\|A\alpha^0 + B(y - F\alpha^0) - \alpha\|^2 \tag{4}$$

over all matrices A and B of appropriate dimensions. Note that the chosen form of the linear estimator $\tilde{\alpha}$ is intuitive and is equivalent to the general form of a linear estimator under the condition of unbiasedness. Also note that the expectation in (4) is with respect to y, ν and also β. Now, the BLUP $\hat{\alpha}$ and its MSE coincide with the LBE $\tilde{\alpha}$ and its Bayes risk, respectively. This follows from the results of Sallas and Harville (1981) and Zehnwirth (1988). Sallas and Harville establish that the BLUP $\hat{\alpha}$ and its MSE can be obtained respectively as limits of BLUPs and MSEs of α defined by the reformulated model (2) as $\sigma_\beta^2 \to \infty$. Zehnwirth (in the context of Kalman filtering) shows that the BLUP of α under model (2) is indeed the LBE and that MSE of BLUP equals the Bayes risk of LBE. Therefore, the LBE $\tilde{\alpha}$ which is the limit of LBEs as $\sigma_\beta^2 \to \infty$ coincides with the BLUP $\hat{\alpha}$ and the same is true of their measures of uncertainty. The corresponding expressions can be obtained as

$$\hat{\alpha} = \tilde{\alpha} = \lim_{\sigma_\beta^2 \to \infty} [\alpha^0 + \Gamma)F^T(F\Gamma)F^T + R)^{-1}(y - F\alpha^0)] \tag{5}$$

and

$$\begin{aligned} \text{MSE}(\hat{\alpha}) &= \text{Bayes Risk } (\tilde{\alpha}) \\ &= \lim_{\sigma_\beta^2 \to \infty} [I - \Gamma F^T(F\Gamma)F^T + R)^{-1}F]\Gamma. \end{aligned} \tag{6}$$

See Sallas and Harville (1981) for closed form expressions of the above limits. An expedient way to get the expressions in (5) and (6) is to think of them respectively as the posterior mean and variance of α under normality. Notice that under normality, the posterior mean is linear and the posterior variance does not depend on y. Therefore, the usual Bayes theory under normality also coincides with the linear Bayes theory when the prior distribution is specified in a wide sense only.

3. CASE OF UNKNOWN VARIANCE COMPONENTS (λ UNKNOWN)

When λ is unknown, it turns out that there is some divergence between the two approaches. It is possible to get some understanding of the differences under normality. Therefore, we assume that the errors ν and ϵ are normal. Also, the number of small areas, m, will be assumed to be large for making asymptotic comparisons. For simplicity, we will illustrate results for the one-fold nested error regression model given by equation (4.2) of Ghosh and Rao, except that we will set $k_{ij} = 1$. Here $\lambda = (\lambda_1, \lambda_2)^T = (\sigma_\nu^2, \sigma^2)^T$ and suppose for illustration that only λ_1 is unknown. The parameters of interest are small area means $\theta_i, i = 1, \ldots m$

where $\theta_i = \bar{X}_i^T\beta + \nu_i$. If λ is known and γ_i denotes $\lambda_1(\lambda_1 + \lambda_2 n_i^{-1})^{-1}$, then the BLUP $\hat{\theta}_i$ and LBE $\tilde{\theta}_i$ (or BUP and BE respectively under normality) are obtained from (5) as

$$\hat{\theta}_i = \tilde{\theta}_i = \bar{X}_i^T\hat{\beta} + \gamma_i(\bar{y}_i - \bar{x}_i^T\hat{\beta}) \tag{7}$$

and from (6); we have, after noting that under normality the Bayes risk is same as the posterior variance (PV),

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_i) &= \mathrm{PV}(\theta_i) \\ &= \lambda_1\lambda_2 n_i^{-1}(\lambda_1 + \lambda_2 n_i^{-1})^{-1} \\ &\quad + (\bar{X}_i - \gamma_i\bar{x}_i)^T(X^TV^{-1}X)^{-1}(\bar{X} - \gamma_i\bar{x}_i) \\ &= g_1(\lambda_1) + g_2(\lambda_1), \text{ say.} \end{aligned} \tag{8}$$

Now, an EBLUP is defined by substituting a consistent estimator $\hat{\lambda}_1$ for λ_1 in $\hat{\theta}_i$ (denote by $\hat{\theta}_i(y, \hat{\lambda}_1)$) and an EB estimator is defined by substituting $\hat{\lambda}_1$ in $\tilde{\theta}_i$, to be denoted by $\tilde{\theta}_i(y, \hat{\lambda}_1)$. For facilitating comparison of the two approaches, we will assume that $\hat{\lambda}_1$ is REML. Clearly, the two estimators so defined are identical. The "naive" approximations to the corresponding measures of uncertainty obtained from (8) by substituting $\hat{\lambda}_1$ for λ_1 are also, of course, identical. The qualifying term "naive" is used to indicate that the extra variability due to estimation of λ_1 is not accounted for.

In the expression (8), the terms $g_1(\lambda_1)$ and $g_2(\lambda_1)$ are respectively $O(1)$ and $O(m^{-1})$. For finding the order of the extra term due to estimation of λ_1, first consider the frequentist approach. It can be shown by the δ-method, similar to equation (5.3) of Ghosh and Rao, that

$$\mathrm{mse}(\hat{\theta}_i(y, \hat{\lambda}_1)) = g_1(\lambda_1) + g_2(\lambda_1) + g_3(\lambda_1) + o(m^{-1}), \tag{9}$$

where $g_3(\lambda_1) = n_i^{-2}\lambda_2^2(\lambda_1 + \lambda_2 n_i^{-1})^{-3}\bar{V}(\hat{\lambda}_1)$ and $\bar{V}(\hat{\lambda}_1)$ is the asymptotic variance of $\hat{\lambda}_1$. Notice that the term $g_3(\lambda_1)$ is also $O(m^{-1})$. Substituting $\hat{\lambda}_1$ in (9), we get an estimate of MSE; but the order of bias is $O(m^{-1})$, not $o(m^{-1})$. This is so because the bias in $g_1(\hat{\lambda}_1)$ is $O(m^{-1})$, although biases in $g_2(\hat{\lambda}_1)$ and $g_3(\hat{\lambda}_1)$ are $o(m^{-1})$. To correct this, the approximation of Prasad and Rao (1990, PR for short) can be used under the assumption $E(\hat{\lambda}_1 - \lambda_1) = o(m^{-1})$ as

$$\begin{aligned} \mathrm{mse}(\hat{\theta}_i(y, \hat{\lambda}_1)) &= [g_1(\hat{\lambda}_1) + g_3(\hat{\lambda}_1)] + g_2(\hat{\lambda}_1) + g_3(\hat{\lambda}_1) \\ &= g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + 2g_3(\hat{\lambda}_1). \end{aligned} \tag{10}$$

For the Bayesian approach, corrections for underestimation of $PV(\theta_i)$ due to estimation of $\hat{\lambda}_1$ can be made by using results of the asymptotic (as $m \to \infty$) hierarchical Bayes (HB) theory (cf. Kass and Steffey, 1989). This technique is justified because the HB estimator (i.e., the posterior mean of θ_i) is asymptotically equivalent to the EB estimator $\tilde{\theta}_i(y, \hat{\lambda}_1)$, the order of error in the approximation being $O(m^{-1})$. Also, the HB technique is convenient in practice because, for large m, the posterior distribution of λ_1 is independent of the choice of prior. Now, analogous to (5.11) of Ghosh and Rao [note that β is absent in the expectation operator because variability due to $\hat{\beta}$ is already accounted for in the $PV(\theta_i)$], we have the posterior variance

$$V(\theta_i|y) = E_{\lambda_1|y}V(\theta_i|y, \lambda_1) + V_{\lambda_1|y}E(\theta_i|y, \lambda_1) \tag{11a}$$

$$= E_{\lambda_1|y}(g_1(\lambda_1) + g_2(\lambda_1)) + V_{\lambda_1|y}\tilde{\theta}_i(y, \lambda_1). \tag{11b}$$

It follows from Kass and Steffey (1989) that

$$E_{\lambda_1|y}(g_1(\lambda_1) + g_2(\lambda)) = g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + O(m^{-1}), \tag{12a}$$

$$\begin{aligned} V_{\lambda_1|y}\tilde{\theta}_i(y, \lambda_1) &= d^2(\hat{\lambda}_1)\bar{V}(\hat{\lambda}_1) + o(m^{-1}) \\ &= g_3^*(\hat{\lambda}_1) + o(m^{-1}), \text{ say,} \end{aligned} \tag{12b}$$

where $d(\hat{\lambda}_1)$ is $(\partial/\partial\lambda_1)\tilde{\theta}_i(y, \lambda_1)|_{\lambda_1=\hat{\lambda}_1}$. Note that if β were known, then $g_3^*(\hat{\lambda}_1)$ simplifies to $\lambda_2^2 n_i^{-2}(\hat{\lambda}_1 + \lambda_2 n_i^{-1})^{-4}(\bar{y}_i - \bar{x}_i^T\beta)^2\bar{V}(\hat{\lambda}_1)$ which is more directly comparable to the term $g_3(\hat{\lambda}_1)$ of the frequentist approximation (9). Incidentally, for β known, $\hat{\lambda}_1$ will be the usual ML and not REML.

In the approximation (12a), the neglected term is $O(m^{-1})$. The accuracy of this approximation can be improved by including terms of order $O(m^{-1})$. By using the δ-method, Singh, Stukel and Pfeffermann (1993) obtain an improved Bayesian approximation as

$$\begin{aligned} V(\theta_i|y) &= [g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + sg_3^{**}(\hat{\lambda}_1)] \\ &\quad + g_3^*(\hat{\lambda}_1) + o(m^{-1}), \end{aligned} \tag{13}$$

where $g_3^{**}(\hat{\lambda}_1)$ is $\lambda_2^2 n_i^{-2}(\hat{\lambda}_1 + \lambda_2 n_i^{-1})^{-2}(\hat{\hat{\lambda}}_1 - \hat{\lambda}_1) - g_3(\hat{\lambda}_1)$. The estimator $\hat{\hat{\lambda}}_1$ denotes an improved (over $\hat{\lambda}_1$) approximation to the posterior mean $E(\lambda_1|y)$ in the sense that $E(\lambda_1|y) = \hat{\hat{\lambda}}_1 + o(m^{-1})$ whereas $E(\lambda_1|y) = \hat{\lambda}_1 + O(m^{-1})$. Note that $\hat{\hat{\lambda}}_1$ can be obtained from the results of Tierney, Kass and Kadane (1989). The expression in (13) is a simplified version of the second order approximation of Kass and Steffey; denote it by KS-II*. Their first order (denote by KS-I) does not include the term $g_3^{**}(\hat{\lambda}_1)$. The approximation KS-II* seems more convenient for term by term comparison with the PR approximation (10) than the original second order KS-II (not considered here).

From (7), (8), (10) and (13), one can compare the Bayesian and frequentist approaches for large m

when λ is unknown. The point estimates are identical or very similar depending on the choice of $\hat{\lambda}_1$ for each approach but the associated measures of uncertainty could be quite different. In addition to the above modifications which rely on the δ-method, Singh, Stukel and Pfeffermann (1993) also obtain a modification of the asymptotic Bayes method of Hamilton (1986) which uses Monte Carlo integration (MCI) for evaluating the two terms of the posterior variance given by (11) thus avoiding computation of partial derivatives. The MCI simply entails generating λ_1-values from the approximate posterior distribution of λ_1 which is given by $N(\hat{\lambda}_1, \overline{V}(\hat{\lambda}_1))$. It is not difficult to show that the order of the neglected terms in the Hamilton (H) approximation is $O(m^{-1})$ and not $o(m^{-1})$. However, if the posterior distribution of λ_1 is approximated by $N(\hat{\hat{\lambda}}_1, \overline{V}(\hat{\lambda}_1))$, then the modified Hamilton (MH) approximation is of the desired order. Singh, Stukel and Pfeffermann (1993) report results of a Monte Carlo study on the frequentist properties of various approximations. Empirically, it is found that the KS-I approximation is biased downward, but KS-II* adds a positive term (similar to PR) and tends to be conservative. The behaviour of the MH approximation is quite similar to KS-II*, but H tends to be more biased downward than KS-I. The performance of the PR approximation is found to be best overall with respect to the frequentist properties, although other approximations provide useful alternatives. In particular, Bayesian approximations KS-II* and MH have the distinct advantage of having a dual interpretation in both frequentist and Bayesian contexts.

It may be noted that if m is quite large, then there will be hardly any difference between various approximations.

4. REMARKS

It is evident from the paper of Ghosh and Rao that great advances have been made in the field of small area estimation by both Bayesians and frequentists. It is also evident from the present discussion that there may be quite a bit of agreement between the two approaches. However, these advanced tools are not in widespread use, especially by statistical agencies conducting large scale complex surveys who face probably the greatest demand for small area statistics. Perhaps, the reason for this is the practitioner's skepticism in modelling complex survey data. Indeed, for complex surveys there is very little by way of model validation and more so for element-level modelling because of possible selection bias [see section 4 of Ghosh and Rao and a recent review by Pfeffermann (1993)]. There is no doubt that the area of model validation for complex survey data needs more research. This is also recognized by Ghosh and Rao and I would like to emphasize by noting that further work in this direction will be a very valuable contribution.

ACKNOWLEDGMENT

This research was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

Comment

Elizabeth A. Stasny

Ghosh and Rao are to be congratulated for their timely paper reviewing methods for small-area estimation. My main complaint is that a paper such as this was not available five years ago when I began working on small-area estimation problems. I particularly enjoyed the historical perspective offered in the demographics methods section of the paper; I was sorry that section was so short since much of the material described in that section is not readily available to statisticians outside of the government agencies.

As the authors noted, there is a growing demand for small-area estimates and a corresponding interest in research on procedures for producing such estimates. The widely publicized debate on adjusting the U.S. population census for the undercount to produce adjusted counts for states and large cities has made many researchers focus on small area estimation problems related to the population census. There are, however, other long-standing small-area estimation programs. One of these is the USDA's program of county-level estimation of crop and live-

Elizabeth A. Stasny is Associate Professor, Department of Statistics, 1958 Neil Avenue, 148D Cockins Hall, Ohio State University, Columbus, Ohio 43210.

stock production. Because most of my work on small-area estimation has been on the problem of producing county-level estimates of crop production, I would like to add a brief discussion of this program to Ghosh and Rao's list of examples. A more detailed description of the National Agricultural Statistics Service (NASS) county estimation program is provided by Iwig (1993).

The USDA's NASS, in cooperation with state governments, has published county crop estimates for every state in every year since 1917. These estimates include acreage, yield and production data for many crops, both common (for example, wheat and corn) and rare (for example, rice and peanuts). For example, the 1990 Annual Report of the Ohio Agricultural Statistics Service includes county-level estimates of number of farms, acres in farms; acres harvested, yield (in bushels per acre), and production (in bushels) for a number of common crops including corn, soybeans, wheat, oats, hay and, for counties producing them, for less common crops such as tomatoes and sugar beets; and grain storage capacity.

Funding for extra data collection and the production of county-level estimates is provided by individual states; the USDA's national Quarterly Agricultural Surveys (QAS) are used to produce estimates only at the state and national level. Although a recent effort has been made to standardize county estimation programs (see the task force report by Bass et al., 1989), the sampling and estimation procedures used in county estimation programs differ from state to state. The task force recommendations for sampling procedures are that each state stratify farms by commodity or group of commodities and by size of operation, choose samples from within these strata, combine these samples into a single sample and delete farms that were already sampled for the QAS (since information for those farms is already available). Within this basic sampling plan, however, individual states have a considerable amount of flexibility. States often choose to sample a high proportion of large farms and of farms that responded to the survey in the previous year (possibly 100% in both cases). In addition, since information on farm operations is required to maintain the control data for the sampling frame, states often sample all farms that have not responded to a survey within a certain number of years.

Typically, 15,000 to 20,000 farms are sampled within a state; the response rate is about 30% with no follow-up of nonrespondents. There is no attempt to obtain sample-based weights for the responding cases since they may have been sampled from one or more commodity lists or from the QAS. Thus, small-area estimation methods that require known sampling weights cannot be used.

Methods of county estimation of crop production vary from state to state. A typical procedure involves initially obtaining the direct county estimates from the data available within each county. (Note that there may be few or no observations within a county for a certain crop, particularly if the county is largely urban or if the crop is relatively rare.) Then one or more experts review the estimates and adjust them in light of their personal knowledge of the farms in the sample, weather conditions, production in the county in previous years and other factors. The experts then look at the implications of the adjustments on the estimated production for the state. These last two steps may be repeated several times until the experts are satisfied with the estimates.

The county estimation program provides an example of the constrained estimation problem described in Section 7.2 of Ghosh and Rao's paper. The QAS data, along with other information such as historical data, administrative data (for example, on land set-aside programs) and weather data, is used by NASS to set the state-level estimates of crop production. Because the QAS is a large, probability sample of farms and the estimates produced using QAS data are believed to be fairly accurate, county estimates are typically constrained to agree with NASS's state estimates. Stasny, Goel and Rumsey (1991) consider the problem of how to scale wheat production estimates to agree with the NASS state total. They consider 1) a constant scaling factor, 2) scaling factors that minimize the sum of squared differences between initial and adjusted estimates and 3) scaling factors that minimize the sum of squared relative differences between initial and adjusted estimates. They found that Method 2) was clearly inferior to the other methods, but there was not much difference between estimates scaled using Methods 1) and 3).

While county estimates obtained following a procedure such as that described above have been used successfully for many years, there are many problems with the procedure. The lack of a formal statistical methodology makes it impossible to repeat the estimation process, to compare estimates from different states and to obtain estimates of the uncertainty. On the other hand, the willingness to base county estimates on expert opinion and on data from many sources (current surveys, historical data and administrative records) makes this an exciting area for continued research.

One area for research is in using information from other crops and from neighboring counties or states to improve the county estimates. Pawel and Fesco (1988) explored historical estimates of crop yields

and found that, as expected, there are high positive correlations between yields in neighboring states and between agronomically related crops that are grown in overlapping regions. This research, however, was conducted at the state rather than county level; it is still an open question whether similar relations will be useful at the county level.

Another area for research is in using the historical data on crop production in current county estimates. A natural way to use this information would be in a Bayesian setting such as the hierarchical Bayes estimates described in Section 5.3 of Ghosh and Rao's paper. Indeed, it seems surprising that a noninformative prior would be used in small-area estimation problems involving census data or data from continuing surveys; there is certainly a wealth of information on which to base an informative prior.

Finally, I would like to mention a success story in research in the production of county estimates. Ghosh and Rao describe the experimental research of Battese, Harter and Fuller (1988) on county estimation of crop production using satellite data. This year, for the first time, Arkansas is using satellite data to aid in production of crop acreage estimates as part of their county estimates program. Over the next few years, other states are expected to begin using such data to aid in the production of their crop acreage estimates.

ACKNOWLEDGMENT

This research was supported in part by the USDA/NASS under Cooperative Agreements 58-3AEU-9-80040 and 43-3AEU-3-80083. The author takes sole responsibility for the work.

Comment

Ib Thomsen

It takes talent and hard work to provide an overview and evaluation of a rapidly evolving subject like small area estimation. In my opinion the authors have succeeded in doing this, and I want to congratulate them with a very useful review. In many statistical offices, substantial methodological work is being done to find suitable estimators for small areas. People involved in such work will be grateful to Ghosh and Rao for their present contribution.

Below I shall communicate some experiences gained when developing and using small area estimates within Statistics Norway. But first a few comments to the example given in Section 6 of the paper. In this example a synthetic population is constructed by fitting a nested error regression model to a business population. For this synthetic population, the EBLUB (or EB) and the HB estimators are shown to produce small area estimators which are superior to the ratio-synthetic and a sample-size dependent estimator. As pointed out by the authors, this demonstrates the advantages of using EBLUB or HB estimators when the model fits the data well. A question remains concerning the robustness of these estimators as compared to the simpler sample-size dependent estimator. A column in Table 3 showing the small area means of the real business population could have thrown some light on the robustness of the estimators studied in the paper.

At Statistics Norway, small area estimators have been used for some years now (Laake, 1978). In the beginning we concentrated on synthetic estimators, but more recently composite estimators are being used. In what follows some of our experiences concerning the feasibility of the EB estimator are presented.

I shall look at a very simple situation in which θ_i, $(i = 1, \ldots, T)$ is a small area parameter, and $\overline{X}_i$, $(i = 1, \ldots, T)$ is a direct estimator such that

$$E(\overline{X}_i|\theta_i) = \theta_i \quad i = 1, \ldots, T.$$

The parameters $\theta_1, \theta_2, \ldots, \theta_T$ are considered realizations of a random variable with unknown distribution $G(\cdot)$. The mean μ and variance σ^2 are assumed to be known or that estimates are available. For a set of small areas, unbiased estimators $\overline{X}_1, \ldots, \overline{X}_T$ are available with conditional distributions equal to the binomial.

When $G(\theta)$ is unknown, empirical Bayes estimators generally employ $(\overline{X}_1, \ldots, \overline{X}_T)$ to estimate $E(\theta|\overline{X}_1, \ldots, \overline{X}_T)$. However, for many distribution, $E(\theta|\overline{X}_1, \ldots, \overline{X}_T)$ cannot be consistently estimated un-

Ib Thomsen is Director of Research, Statistics Norway, and Professor of Statistics, University of Oslo, P.B. 8131 Dep, 0033 Oslo, Norway.

less other assumptions are made. Therefore, one often restricts attention to linear estimators, $c = a\overline{X} + b$. Within this class, the estimator which minimizes the mean squared error depends only upon the first two prior moments, both of which can often be estimated with $(\overline{X}_1, \ldots, \overline{X}_T)$. The optimal linear estimator is often the same as the unrestricted Bayes estimator derived under a conjugate prior (Rao, 1976). When the conditional distribution of $\overline{X}_i$ is binomial, the optimal linear estimator is a composite estimator,

$$c_i = W_i\overline{X}_i + (1 - W_i)\mu ,$$

where

$$W_i = \sigma^2\{(1 - 1/n_i)\sigma^2 + \mu(1 - \mu)/n_i\}^{-1}$$

and n_i denotes the number of observations from small area i (Spjøtvoll and Thomsen, 1987). With these weights we have that

$$(1)\quad E\left\{(1/T)\sum_{i=1}^{T}(c_i - \mu)^2\right\} = \sigma^2(1/T)\sum_{i=1}^{T}W_i \le \sigma^2.$$

It follows that the variation between the small area estimators can be much smaller than the prior known variance. I have often observed this phenomenon in practice; a consequence is usually that the range of the small area estimators is much smaller than expected. (Expectations are based on information outside the sample.) In practice the parameter σ^2 is often of great importance in itself. As said in the introduction, "Increasing concern with issues of distribution, equity and disparity (Brackstone, 1987)." To me, this means that the disparity between the small area is important and should be easily read from a table presenting small-area estimators. As mentioned by Ghosh and Rao, there are composite estimators which have the same expectation and variance as the prior distribution, one of which is simply to use $\{W_i\}^{1/2}$ instead of W_i as weights in the composite estimator.

When area-specific auxiliary information is available and a model like (4.1) in the paper is used, I have often observed a similar "overshrinkage" as under the simpler model above. An inequality similar to (1) can be found under model (4.1), but now σ^2 denotes the variance of the residual in equation (4.1). Again $\{W_i\}^{1/2}$ can be used to avoid "overshrinkage".

Due to the often observed "overshrinkage" and the fact that our models seem too complicated to many of our users of small-area estimators, I have often found it very difficult to make them use the optimal estimators presented in the paper. On the other hand, a number of sample-size dependent estimators are more easily "sold" to the user and therefore more used up until now.

In Statistics Norway a number of administrative registers are available and used to construct small-area estimators. In many cases it is natural to use nested error regression models. However, progress in this area has been slow due to difficulties concerning model diagnotics for linear models involving random effects. I therefore find Section 7.1 particularly interesting and shall use this section intensively in our further hunt for feasible small area estimates.

Rejoinder

M. Ghosh and J. N. K. Rao

We thank the discussants for their insightful comments as well as for providing various extensions of the models and the methods reviewed in our paper. These expert commentaries have brought out many diverse issues and concerns related to small area estimation, particularly on the model-based methods.

Several discussants emphasised the importance of model diagnostics in the context of small area estimation. We agree wholeheartedly with the discussants on this issue. As noted in Section 7.1 of our article, the literature on this topic is not extensive, unlike standard regression diagnostics. We hope that future research on small area estimation will give greater emphasis to model validation issues.

A second concern expressed by some of the discussants is that the composite estimators typically used for small area estimation may "overshrink" towards a synthetic estimator. Thomsen, in his discussion, suggests that a larger weight should be given to the direct estimator. We agree with his suggestion but are hesitant to recommend blanket use of the weight $W_i^{1/2}$, instead of W_i, to the direct estimator $(0 < W_i < 1)$. We believe that the weight should be determined adaptively meeting certain optimality criteria as in Louis (1984) and Ghosh (1992). Cressie and Kaiser, in their discussion, address con-

strained estimation at some length, emphasising the multivariate aspects of the problem but not invoking any optimality conditions.

Cressie and Kaiser as well as Holt suggest possible extensions of the two basic small area models (4.4) and (4.6) given in Section 4. Their general hierarchical modeling ensures that the variability among observation vectors for the different small areas is attributable not only to sampling variability but also to variability among the associated regression coefficients, β_i. Holt's model looks promising since it allows the β_i to depend on area level auxiliary variables, Z_i, thus effectively integrating the use of unit level and area level covariates into a single model.

A slightly less general version of Cressie and Kaiser's hierarchical model (3) appears in Datta and Ghosh (1991) where a full hierarchical Bayes analysis is presented. In an earlier version of our paper (Ghosh and Rao, 1991) we have in fact considered the general model of Datta and Ghosh but decided to abandon it in the revision in favour of the simpler, but widely used, models (4.4) and (4.5) in order to keep the discussion more accessible to a general readership and the notation simple.

We now turn to some of the specific points raised by the discussants.

CRESSIE AND KAISER

Cressie and Kaiser stress the importance of nonlinear modelling which is especially needed for binary and count data. Our Section 7.3 gives a brief account of logistic regression and log-linear models suitable for such data. These can be viewed as special cases of generalized linear models (McCullagh and Nelder, 1989). Zerger and Karim (1991) have studied generalized linear models with random effects using a Gibbs sampling approach. Their results may be applicable to small area estimation. In a 1993 Ph.D. thesis at the University of Florida, Kannan Natarajan implemented an extensive hierarchical Bayes analysis under generalized linear models in the context of two-stage sampling within small areas. He used the Metropolis within Gibbs sampling algorithm (cf. Müller, 1991). His method is easier to implement than the procedure of Zeger and Karim (1991) due to logconcavity of certain posterior distributions which permits the use of adaptive rejection sampling of Gilks and Wild (1992).

We agree with Cressie and Kaiser regarding the multivariate aspects of small area estimation. Our analysis can be extended to produce approximately unbiased estimators of the off-diagonal elements of the mean-square error matrix as well as to obtain exact posterior covariances of small area means. Reporting these quantities in tables, however, is usually cumbersome since there will be $\binom{m}{2}$ such quantities when the number of small areas is m. Nevertheless, these estimates should be available, as they are needed in calculating measures of uncertainty at higher levels of aggregation.

HOLT

The example in Section 6 of our paper, based on a simple random sample drawn from a synthetic population, was introduced mainly to illustrate the proposed methods. We agree with Holt that a simulation study based on repeated samples from the population is better for comparing the relative performances of estimators. Such a simulation study was, in fact, conducted by Choudhry and Rao (1993) using both real and synthetic populations. Comparisons were made under customary repeated sampling (approach (c) of Holt) as well as under a conditional framework by fixing the values of samples sizes, n_i (approach (b) of Holt).

We also agree with Holt that one should be cautious in comparing relative performances based on summary measures, obtained by averaging across all small areas, without paying some attention to the distribution. Such summary measures, however, may be quite useful in an overall comparison of competing estimators, especially when there is no clear-cut winner when the small areas are judged individually.

SCHAIBLE AND CASADY

Despite many success stories of model-based indirect estimators, there are some practical problems associated with their use. We are grateful to Schaible and Casady for providing a comprehensive list of such problems.

We agree with them that models based on expediency "instill little confidence in either the producers or consumers of the estimators." Model diagnostics should be an integral part of any model-based procedure in order to alleviate this problem.

SINGH

We are glad that Singh has investigated under a simplified model some frequentist properties of the Kass-Steffey first-order approximation (KS-I) to the posterior variance and Hamilton's (1986) Monte Carlo integration method (H) of evaluating the posterior variance. He also suggests modifications, KS-II* and MH, to improve their accuracy; in particular his formula (13) which is a simplified version of the second-order approximation of Kass and Steffey. It

would be useful to provide similar improved approximations for more complex models and to study their frequentist properties.

We agree with Singh that the Bayesian approximations KS-II* and MH have the advantage of dual interpretation in both frequentist and Bayesian contexts, although Prasad Rao's estimator of MSE performed better with respect to frequentist properties.

In the case of known variance components, Singh has demonstrated that the linear Bayes estimate (LBE) and its Bayes risk coincide with the BLUP estimator and its MSE. A similar result appears in Datta et al. (1992).

STASNY

Stasny provides an excellent account of USDA's program of country-level estimation of crop and livestock production. She also raises the important issue that the current small area estimation methods need to be modified in the presence of nonresponse. In this regard, Stasny's (1991) important work on hierarchical models for the probabilities of a survey classification and nonresponse might be relevant. We might add that the role of measurement error in small domain estimation is also important. Eltinge and Harter (1990) have studied the effect of measurement errors and propose some modified small area estimators.

We agree with Stasny that in some small area estimation problems historical data can be used to construct informative priors and obtain the resulting hierarchical Bayes estimates.

We are also delighted to learn about the success story that Arkansas is currently using satellite data, in conjunction with USDA survey data, for the production of county estimates based on small area models.

THOMSEN

We agree with Thomsen's observation that many users find the small area models too complicated and are bothered by the overshrinkage problem associated with the optimal estimators. Further work on model diagnostics and constrained estimation and the development of suitable packages to implement both model selection and estimation should alleviate this problem.

Thomsen also remarks that sample-size dependent estimators, such as those based on the weights (3.6), are more easily "sold" to the user. Such estimators are clearly useful and computationally attractive, but their limitations should also be noted. As mentioned in Section 3 of our paper, sample-size dependent estimators can fail to borrow strength from related domains even when the expected domain sample size, $E(n_i)$, is not large enough to make the direct estimators reliable. These estimators were originally designed to handle domains for which $E(n_i)$ is large enough to make the direct estimators satisfy reliability requirements (Drew, Singh and Choudhry, 1982). Another disadvantage of sample-size dependent estimators, noted in Section 3, is that the weights do not take account of the size of between area variation relative to within area variation for the characteristic of interest, unlike model-based estimators. Choudhry and Rao (1993) demonstrate that large efficiency gains can be achieved by using the EBLUP estimators when the between area variation is small relative to within area variation.

ADDITIONAL REFERENCES

ALBERT, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83** 1037–1044.

ALBERT, J. H. and PEPPLE, P. A. (1989). A Bayesian approach to some overdispersion models. *Canad. J. Statist.* **17** 333–344.

BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., and WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates, USDA National Agricultural Statistics Service, Washington, DC.

CHOUDHRY, G. H. and RAO, J. N. K. (1993). Evaluation of small area estimators: an empirical study. In *Small Area Statistic and Survey Designs* **1** 271–290. Central Statistical Office, Warsaw.

CRESSIE, N. (1986). Empirical Bayes estimation of undercount in the decennial census. Statistical Laboratory Preprint 86–58. Iowa State Univ., Ames.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology* **14** 191–208.

CRESSIE, N. (1990b). Weighted smoothing of estimated undercount (with discussion). In *Proceedings of Bureau of the Census 1990 Annual Research Conference* 301–325, 362–366. U.S. Bureau of the Census, Washington, DC.

CRESSIE, N. (1992). Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis* **24** 75–95.

ELTINGE, J. L. and HARTER, R. L. (1990). Small domain estimation in the presence of measurement and sampling errors. Technical Report, Dept. Statistics, Texas A&M Univ.

GHOSH, M. and RAO, J. N. K. (1991). Small area estimation: an appraisal. Technical Report 390, Dept. Statistics, Univ. Florida, Gainesville.

GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Roy. Statist. Soc. Ser. C* **41** 337–348.

GOEL, P. K. and DEGROOT, M. H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76** 140–147.

HAMILTON, J. D. (1986). A standard error for the estimated state vector of a state-space model. *J. Econometrics* **33** 387–397.

HOLT, D. and MOURA, F. (1993). Mixed models for making small area estimates. In *Small Area Statistics and Surveys Designs* (G. Kalton, J. Kordos and R. Platek, eds.) **1** 221–231. Central Statistical Office, Warsaw.

IWIG, W. C. (1993). "The National Agricultural Statistics Service County Estimates Program", in "Indirect Estimators in Federal Programs", Statistical Policy Working Paper 21, Report

of the Federal Committee on statistical Methodology, Subcommittee on Small Area Estimation, Washington, DC, 7.1-7.15.

LAAKE, P (1978). An evaluation of synthetic estimates of employment. *Scand. J. Statist.* 5 57–60.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* 34 1–41.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York.

MÜLLER, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report 91–09, Dept. Statistics, Purdue Univ.

PAWEL, D. and FESCO, R. (1988). On the use of correlations in crop yields. *Proceedings of the Section on Survey Research Methodology* 391–396. Amer. Statist. Assoc., Washington, DC.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Internat. Statist. Rev.* 61 317–337.

RAO, C. R. (1976). Characterization for prior distributions and solutions to a compound decision problem. *Ann. Statist.* 4 823–835.

SALLAS, W. M. and HARVILLE, D. A. (1981). Best linear recursive estimation for mixed linear models. *J. Amer. Statist. Assoc.* 76 860–869.

SINGH, A. C., STUKEL, D. M. and PFEFFERMANN, D. (1993). Bayesian versus frequentist measures of uncertainty for small area estimators. *Proceedings of the Section on Survey Research Methods*, Amer. Statist. Assoc. To appear.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* 55 3–23, 53–102.

STASNY, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: an example from the National Crime Survey. *J. Amer. Statist. Assoc.* 86 296–303.

STASNY, E. A., GOEL, P. K. and RUMSEY, D. J. (1991). County estimates of wheat production. *Survey Methodology* 17 211–225.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* 84 710–716.

TUKEY, J. W. (1974). Named and faceless values: An initial exploration in memory of Prasanta C. Mahalanobis. *Sankhyā Ser. A* 36 125–176.

TUKEY, J. W. (1983). Affidavit presented to District Court, Southern District of New York. *Cuomo et al. versus Baldridge.* 80 Civ. 4550 (JES).

WOLTER, K. M. and CAUSEY, B. D. (1991). Evaluation of procedures for improving population estimates for small areas. *J. Amer. Statist. Assoc.* 86 278–284.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* 86 79–86.

ZEHNWIRTH, B. (1988). A generalization of the Kalman filter for models with state-dependent observation variance. *J. Amer. Statist. Assoc.* 83 164–167.

Generalized Linear Models for Small-Area Estimation

Malay GHOSH, Kannan NATARAJAN, T. W. F. STROUD, and Bradley P. CARLIN

Bayesian methods have been used quite extensively in recent years for solving small-area estimation problems. Particularly effective in this regard has been the hierarchical or empirical Bayes approach, which is especially suitable for a systematic connection of local areas through models. However, the development to date has mainly concentrated on continuous-valued variates. Often the survey data are discrete or categorical, so that hierarchical or empirical Bayes techniques designed for continuous variates are inappropriate. This article considers hierarchical Bayes generalized linear models for a unified analysis of both discrete and continuous data. A general theorem is provided that ensures the propriety of posteriors under diffuse priors. This result is then extended to the case of spatial generalized linear models. The hierarchical Bayes procedure is implemented via Markov chain Monte Carlo integration techniques. Two examples (one featuring spatial correlation structure) are given to illustrate the general method.

KEY WORDS: Hierarchical model; Markov chain Monte Carlo; Posterior propriety; Spatial statistics.

1. INTRODUCTION

Bayesian methods have been used quite extensively in recent years for solving small-area estimation problems. Particularly effective in this regard have been the hierarchical Bayes (HB) and empirical Bayes (EB) approaches, which are especially suitable for a systematic connection of local areas through the model. For the general theory as well as specific applications of the HB and EB methods for small-area estimation, relevant work includes that of Datta and Ghosh (1991), Fay and Herriot (1979), Ghosh and Lahiri (1987, 1992), Ghosh and Meeden (1986), Prasad and Rao (1990), and Stroud (1987, 1991), among others. Ghosh and Rao (1994) have provided a review of many of these results.

But development to date has concentrated mainly on continuous-valued variates. Often the survey data are discrete or categorical, for which the HB or EB analysis suitable for continuous variates is not appropriate. Recently, some work has begun to appear on the Bayesian analysis of binary survey data. Dempster and Tomberlin (1980), Farrell, MacGibbon, and Tomberlin (1997) and MacGibbon and Tomberlin (1989) have obtained small area estimates of proportions via EB techniques, whereas Malec, Sedransk, and Tompkins (1993) found the predictive distributions of a linear combination of binary random variables using a HB technique. Stroud (1991) also developed a general HB methodology for binary data, and Nandram and Sedransk (1993) suggested Bayesian predictive inference for binary data from a two-stage cluster sample. Subsequently, Stroud (1994) provided a comprehensive treatment of binary survey data encompassing simple random, stratified, cluster, and two-stage sampling, as well as two-stage sampling within strata.

The binary models constitute a subclass of generalized linear models that are often used for a unified analysis of both discrete and continuous data. Section 2 presents a general account of how HB generalized linear models (GLMs) can be used for small-area estimation. The section begins with a general description of HB GLMs. Sufficient conditions are provided for the joint posterior distribution of the parameters of interest to be proper under the proposed hierarchical models. The Bayes procedure is implemented via Markov chain Monte Carlo (MCMC) integration techniques—in particular, using the Gibbs sampler. Next, this section contains a discussion of some general multicategory models that may be handled indirectly by methods of this section, even though in their natural multinomial formulation they do not fit into the univariate GLM framework. We also point out that in contrast to the work of Stroud (1994), who used the Brooks (1984) method for approximating numerical integrals, we use exact MCMC integration techniques. We conclude this section by considering some spatial GLMs and find sufficient conditions that ensure the propriety of the posterior. We also point out a common HB model for this situation that actually leads to an improper posterior.

Section 3 contains the analysis of two real datasets. The first consists of responses to the question "Have you experienced any negative impact of exposure to health hazards in the workplace?" based on a 1991 sample of all persons in 15 geographic regions of Canada (Statistics Canada 1992). For each region, workers were classified by age (≤ 40 or >40) and sex (male or female). The responses were classified into four categories: (1) yes, (2) no, (3) not exposed, and (4) not applicable or not stated. The objective is to estimate the proportion of workers in each of the four categories for every one of the $15 \times 2 \times 2 = 60$ groups cross-classified by 15 geographic regions and the 4 demographic categories.

Malay Ghosh is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Kannan Natarajan is Senior Research Biostatistician, Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ 08543. T. W. F. Stroud is Professor, Department of Mathematics and Statistics, Queens University, Kingston, ON K7L 3N6, Canada. Bradley P. Carlin is Associate Professor, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455. This research was supported in part by National Science Foundation grants SES-9201210 and SBR-9423996 (Ghosh), a grant from the Natural Science and Engineering Research Council of Canada (Stroud), and National Institute of Allergy and Infectious Diseases FIRST Award 1-R29-AI33466 and National Institute of Environmental Health Sciences grant 1-R01-ES07750 (Carlin). The authors thank James H. Albert of Bowling Green State University for many useful conversations, Robert Tsutakawa for his diligence in tracking down the county identifiers for the Missouri lung cancer dataset, and an associate editor and three referees for their thoughtful remarks, which led to substantial improvements in this article.

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

Our HB cell probability estimates "borrow strength" from the other cells, resulting in smaller standard errors. Moreover, shrinkage toward the grand mean is done adaptively, in that the estimates reported for cells with larger sample sizes are shrunk less than those based on smaller sample sizes.

The second dataset relates to cancer mortality rates for the 115 counties in Missouri during 1972–1981. In each county, deaths due to lung cancer are broken down into four age groups (45–54, 55–64, 65–74, and 75+) and two sex groups (male and female). The number of deaths in some of these county subgroups during this period is very small (occasionally 0), so there is a clear need to borrow strength across cells. Tsutakawa (1988) and Tsutakawa, Shoop, and Marienfeld (1985) considered EB estimation of the rates for the given age groups, and Tsutakawa (1985) compared these EB rates with approximate Bayes rates, but these works dealt only with the male population and did not use prior distributions that could account for spatial similarity of the underlying rates in neighboring counties. We consider several possible models using such a spatial smoothing prior and including age, sex, and age–sex interaction as covariates. After selecting an appropriate model somewhat informally using a log-likelihood score statistic, we map the raw and fitted relative risks for a particular age–sex group as well as the fitted risks obtained in the earlier analysis by Tsutakawa, allowing the benefits of our spatial model to be assessed visually. We also investigate the adequacy of our model using a variety of model checks facilitated by our MCMC implementation.

2. HIERARCHICAL MODELS

Suppose that there are m strata or local areas. Let Y_{ik} denote the minimal sufficient statistic (discrete or continuous) for the kth unit within the ith stratum ($k = 1, \ldots, n_i; i = 1, \ldots, m$). The Y_{ik} are assumed to be conditionally independent with pdf

$$f(y_{ik}|\theta_{ik}, \phi_{ik}) = \exp[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik})) + \rho(y_{ik}; \phi_{ik})] \tag{1}$$

($k = 1, \ldots, n_i; i = 1, \ldots, m$). Such a model is referred to as a generalized linear model (McCullagh and Nelder 1989, p. 28). The density (1) is parameterized with respect to the canonical parameters θ_{ik} and the scale parameters ϕ_{ik} (> 0). It is assumed that the scale parameters ϕ_{ik} are known.

The natural parameters θ_{ik} are first modeled as

$$h(\theta_{ik}) = \mathbf{x}_{ik}^T\beta + u_i + \varepsilon_{ik} \qquad (k = 1, \ldots, n_i; i = 1, \ldots, m), \tag{2}$$

where h is a strictly increasing function, the $\mathbf{x}_{ik}$ ($p \times 1$) are known design vectors, β ($p \times 1$) is the unknown regression coefficient, u_i are the random effects, and ε_{ik} are the errors. It is assumed that the u_i and the ε_{ik} are mutually independent with $u_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_u^2)$ and $\varepsilon_{ik} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

It is possible to represent (1) and (2) in a hierarchical framework. Let $R_u = \sigma_u^{-2}$ and $R = \sigma^{-2}$. Also, let $\theta = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{m1}, \ldots, \theta_{mn_m})^T$ and $\mathrm{u} = (u_1, \ldots, u_m)^T$. Then the hierarchical model is given by the following:

(I) Conditional on $\theta, \beta, \mathrm{u}, R_u = r_u$, and $R = r, Y_{ik}$ are independent with densities given in (1).

(II) Conditional on $\beta, \mathrm{u}, R_u = r_u$, and $R = r, h(\theta_{ik}) \overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_{ik}^T\beta + u_i, r^{-1})$.

(III) Conditional on $\beta, R_u = r_u$, and $R = r, u_i \overset{\text{ind}}{\sim} \mathrm{N}(0, r_u^{-1})$.

To complete the hierarchical model, we assign the following prior to $\beta, R_u = r_u$, and $R = r$:

(IV) $\beta, R_u = r_u$, and $R = r$ are mutually independent with $\beta \sim \text{uniform}(\mathrm{R}^p), (p < m), R_u \sim \text{gamma}(\frac{1}{2}a, \frac{1}{2}b)$, and $R \sim \text{gamma}(\frac{1}{2}c, \frac{1}{2}d)$.

(A random variable $Z \sim \text{gamma}(\alpha, \beta)$ if Z has pdf $f(z) \propto \exp(-\alpha z)z^{\beta-1}I_{(0,\infty)}(z)$.)

We are interested in finding the joint posterior distribution of the $g(\theta_{ik})$'s, where g is a strictly increasing function, given the data $\mathrm{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m})^T$, and in particular in finding the posterior means, variances, and covariances of these parameters. In typical applications, $g(\theta_{ik}) = \psi'(\theta_{ik}) = E(Y_{ik}|\theta_{ik})$.

First, however, one must ensure that the joint posterior distribution of the θ_{ik}'s given y is proper. A theorem is proved to this effect. In what follows, the support of θ_{ik} is the open interval $(\underline{\theta}_{ik}, \bar{\theta}_{ik})$, where the lower endpoint of the interval can be $-\infty$, the upper endpoint can be $+\infty$, or both.

Theorem 1. Assume that $a > 0, c > 0, \sum_i n_i - p + d > 0$, and $m + b > 0$. Then, if

$$\int_{\underline{\theta}_{ik}}^{\bar{\theta}_{ik}} \exp\{[\theta y_{ik} - \psi(\theta)]/\phi_{ik}\}h'(\theta)\, d\theta < \infty \tag{3}$$

for all y_{ik} and ϕ_{ik} (> 0), the joint posterior pdf of the θ_{ik}'s given y is proper.

The proof of this theorem is deferred to the Appendix.

Two special cases are of interest. In the first case, $Y_{ik}|\theta_{ik} \sim \text{bin}\,(n_{ik}, \exp(\theta_{ik})/(1 + \exp(\theta_{ik})))$. Suppose now that h is the identity function; that is, the link is canonical. Also, let $g(\theta_{ik}) = \psi'(\theta_{ik})/n_{ik} = \exp(\theta_{ik})/[1 + \exp(\theta_{ik})]$. Then, writing $p_{ik} = \exp(\theta_{ik})/[1 + \exp(\theta_{ik})]$, (3) reduces to $\int_0^1 p_{ik}^{y_{ik}-1}(1 - p_{ik})^{n-y_{ik}-1}\, dp_{ik} < \infty$, which requires $1 \le y_{ik} \le (n_{ik} - 1)$; that is, excludes cases of all failures or all successes. In the second case, $Y_{ik}|\theta_{ik} \sim \text{Poisson}(\exp(\theta_{ik}))$. Then, if h is the canonical link, and $g(\theta_{ik}) = \psi'(\theta_{ik}) = \exp(\theta_{ik})$, (3) reduces to $\int_0^\infty \zeta_{ik}^{y_{ik}-1}\exp(-\zeta_{ik})\, d\zeta_{ik} < \infty$, which holds for $y_{ik} = 1, 2, \ldots$ It may be noted, however, that although our general theorem needs this postivity restriction on the y_{ik} in the binomial and Poisson examples, recent work by Maiti (1997) showed that $\sum_k y_{ik} > 0$ for each i is sufficient for posterior propriety.

Direct evaluation of the joint posterior distribution of the $g(\theta_{ik})$'s given y involves high-dimensional numerical integration and is not computationally feasible. Instead, we use

the Gibbs sampler (Gelfand and Smith 1990; Geman and Geman 1984). Its implementation requires generating samples from certain conditional posterior distributions. Write $\mathrm{h}(\theta) = (h(\theta_{11}), \ldots, h(\theta_{1n_1}), \ldots, h(\theta_{m1}), \ldots, h(\theta_{mn_m}))^T$, and $\mathrm{X} = (\mathrm{x}_{11}, \ldots, \mathrm{x}_{1n_1}, \ldots, \mathrm{x}_{m1}, \ldots, \mathrm{x}_{mn_m})^T$. Assume that $\mathrm{X}^T\mathrm{X}$ is nonsingular. The necessary conditional distributions based on (I)–(IV) are

(i) $\beta|\theta, \mathrm{u}, r_u, r, \mathrm{y} \sim N((\mathrm{X}^T\mathrm{X})^{-1}(\mathrm{X}^T\mathrm{h}(\theta) - \sum_i u_i \sum_k \mathrm{x}_{ik}), r^{-1}(\mathrm{X}^T\mathrm{X})^{-1})$;

(ii) $u_i|\theta, \beta, r_u, r, \mathrm{y} \overset{\text{ind}}{\sim} N((rn_i + r_u)^{-1} r \sum_k (h(\theta_{ik}) - \mathrm{x}_{ik}^T\beta), (rn_i + r_u)^{-1})$;

(iii) $R|\theta, \beta, \mathrm{u}, r_u, \mathrm{y} \sim \text{gamma}(\frac{1}{2}(c + \sum_i \sum_k (h(\theta_{ik}) - \mathrm{x}_{ik}^T\beta - u_i)^2), \frac{1}{2}(d + \sum_1^m n_i))$;

(iv) $R_u|\theta, \beta, \mathrm{u}, r, \mathrm{y} \sim \text{gamma}(\frac{1}{2}(a + \sum_i u_i^2), \frac{1}{2}(b + \sum_1^m n_i))$; and

(v) $\theta_{ik}|\beta, \mathrm{u}, r_u, r, \mathrm{y} \overset{\text{ind}}{\sim} \pi(\theta_{ik}|\beta, \mathrm{u}, r_u, r, \mathrm{y})$

$$\propto \exp\Big[(y_{ik}\theta_{ik} - \psi(\theta_{ik}))\phi_{ik}^{-1} - \frac{r}{2}(h(\theta_{ik}) - \mathrm{x}_{ik}^T\beta - u_i)^2\Big] h'(\theta_{ik}).$$

It is easy to generate samples from the normal and gamma distributions given in (i)–(iv). On the other hand, as evidenced in (v), the posterior distribution of θ_{ik} given $\beta, \mathrm{u}, r_u, r$, and y is known only up to a multiplicative constant, and accordingly one must use a general accept–reject algorithm to generate samples from this pdf. In the special case where h is the identity function, the task becomes much simpler due to the following lemma, which establishes log-concavity of $\pi(\theta_{ik}|\beta, \mathrm{u}, r_u, r, \mathrm{y})$. In such cases one can use the adaptive rejection sampling scheme of Gilks and Wild (1992).

Lemma 1. When $h(z) = z$ for all z, $\log \pi(\theta_{ik}|\beta, \mathrm{u}, r, r_u, \mathrm{y})$ is a concave function of θ_{ik}.

Proof. Straightforward.

Inference about θ will be based on (i)–(v). Indeed, based on (v), one can also find $E(\theta_{ik}|\mathrm{y}), V(\theta_{ik}|\mathrm{y})$, and $\mathrm{cov}(\theta_{ik}, \theta_{i'k'}|\mathrm{y})(i,k) \neq (i',k')$ based on Monte Carlo integration techniques and formulas for iterated conditional expectations and variances.

The model considered in (I)–(IV) resembles closely the ones considered by Breslow and Clayton (1993), MacGibbon and Tomberlin (1989), and Zeger and Karim (1991). However, this model is not strictly contained in the one considered by Zeger and Karim (1991). Zeger and Karim considered $h(\theta_{ik}) = \mathrm{x}_{ik}^T\beta + u_i$, where $h(\cdot)$ is a strictly increasing function, but this formulation does not include possible error in misspecifying this model. Indeed, according to our model, the uncertainty in specifying the model is broken up into two components: the effect of the local area and the error component. This allows the possibility of accounting for overdispersion by introducing an extra variance component.

Our method should also be contrasted to that of Albert (1988), which generalizes the approach of Leonard and Novick (1986) and which was applied to binary survey data by Stroud (1994). Albert's method applied to the present setting first uses independent conjugate priors

$$\pi(\theta_{ik}|m_{ik}, \zeta) = \exp[\zeta(m_{ik}\theta_{ik} - \psi(\theta_{ik})) + g(m_{ik}, \zeta)] \quad (4)$$

for the θ_{ik}. Next, it assumes that $h(m_{ik}) = \mathrm{x}_{ik}^T\beta$ for some known monotone function h. Subsequently, it assigns distributions (possibly diffuse) to the hyperparameters β and ζ. In contrast, our approach does not need the conjugacy of the prior and models monotone functions of θ_{ik} instead of monotone functions of $m_{ik} = E[\psi'(\theta_{ik})]$. Moreover, Albert (1988) suggested approximation to the Bayes procedure by one of the following three methods: Laplace's method, quasi-likelihood approaches, or Brooks's (1984) method. These approximations generally are unnecessary now with the advent of the sophisticated MCMC integration techniques.

The log-concavity idea is used slightly differently by Dellaportas and Smith (1993), whose prime objective is inference about β in generalized linear models and model θ_{ik} as functions of β without any error. In addition, their method, unlike ours, does not introduce any uncertainty in specifying the model.

We now examine how the previous results can be generalized for the analysis of multicategory data. Consider m strata labeled $1, \ldots, m$. Within each stratum, several units are selected; suppose that the responses of individuals within each selected unit are independent and can be classified into J categories. For the kth selected unit within the ith stratum, let p_{ijk} denote the probability that an individual's response falls in the jth category ($j = 1, \ldots, J; k = 1, \ldots, n_i$). Then within the kth selected unit within the ith stratum, Z_{ijk} ($j = 1, \ldots, J$) have a joint multinomial$(t_{ik}; p_{i1k}, \ldots, p_{iJk})$ distribution, where $t_{ik} = \sum_j Z_{ijk}$. Using the well-known relationship between the multinomial and Poisson distributions, $(Z_{i1k}, \ldots, Z_{iJk})$ has the same distribution as the joint conditional distribution of $(Y_{i1k}, \ldots, Y_{iJk})$ given $\sum_{j=1}^J Y_{ijk} = t_{ik}$, where the Y_{ijk} ($j = 1, \ldots, J$) are independent Poisson(ζ_{ijk}) and $p_{ijk} = \zeta_{ijk} / \sum_{j=1}^J \zeta_{ijk}$ ($j = 1, \ldots, J$).

Let $\theta_{ijk} = \log \zeta_{ijk}$, and let θ denote the vector whose elements are the θ_{ijk}'s. One can also model θ_{ijk} as

$$h(\theta_{ijk}) = \mathrm{x}_{ijk}^T\beta + u_{ij} + \varepsilon_{ijk}. \quad (5)$$

Also, it is assumed that u_{ij} and the ε_{ijk} are mutually independent with $u_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_u^2)$ and $\varepsilon_{ijk} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

Then the hierarchical model, which is closely related to (I)–(IV), is given by the following:

(A) $Y_{ijk}|\theta, \mathrm{u}, \beta, r_u, r$ are independent with

$$f(y_{ijk}|\theta, \mathrm{u}, \beta, r_u, r) = \exp[\phi_{ijk}^{-1}(y_{ijk}\theta_{ijk} - \psi(\theta_{ijk})) + \rho(y_{ijk}; \phi_{ijk})].$$

(B) $h(\theta_{ijk})|\mathrm{u}, \beta, r_u, r \overset{\text{ind}}{\sim} \mathrm{N}(\mathrm{x}_{ijk}^T\beta + u_{ij}, r^{-1})$.

(C) $u_{ij}|\beta, r_u, r \overset{\text{ind}}{\sim} \mathrm{N}(0, r_u^{-1})$.

(D) β, R_u, and R are mutually independent with $\beta \sim \text{uniform}(\mathrm{R}^p)$, $R_u \sim \text{gamma}(\frac{1}{2}a, \frac{1}{2}b)$, and $R \sim \text{gamma}(\frac{1}{2}c, \frac{1}{2}d)$.

We are interested in the posterior means, variances, and covariances of the $p_{ijk} = \exp(\theta_{ijk})/\sum_{j=1}^{J}\exp(\theta_{ijk})$ $(k = 1,\ldots,n_i; i = 1,\ldots,m; j = 1,\ldots,J)$. The necessary posterior distributions for doing these calculations are given by

(a) $\beta|\theta, \mathbf{u}, r_u, r, \mathbf{y} \sim N((\sum_{i,j,k}\mathbf{x}_{ijk}\mathbf{x}_{ijk}^T)^{-1}(\sum_{i,j,k}\mathbf{x}_{ijk}(h(\theta_{ijk}) - u_{ij})), r^{-1}(\sum_{i,j,k}\mathbf{x}_{ijk}\mathbf{x}_{ijk}^T)^{-1})$;

(b) $u_{ij}|\theta,\beta,r_u,r,\mathbf{y} \overset{\text{ind}}{\sim} \mathrm{N}((rn_i + r_u)^{-1}r\sum_k(h(\theta_{ijk}) - \mathbf{x}_{ijk}^T\beta), (rn_i + r_u)^{-1})$;

(c) $R|\theta,\beta,\mathbf{u},r_u,\mathbf{y} \sim \text{gamma}(1/2(c + \sum_{i,j,k}(h(\theta_{ijk}) - \mathbf{x}_{ijk}^T\beta - u_{ij})^2), 1/2(d + J\sum_i n_i))$;

(d) $R_u|\theta,\beta,\mathbf{u},r,\mathbf{y} \sim \text{gamma}(1/2(a + \sum_i\sum_j u_{ij}^2), 1/2(b + mJ))$; and

(e) $\theta_{ijk}|\beta,\mathbf{u},r_u,r,\mathbf{y} \overset{\text{ind}}{\sim} \pi(\theta_{ijk}|\beta,\mathbf{u},r_u,r,\mathbf{y}) \propto \exp[(y_{ijk}\theta_{ijk} - \psi(\theta_{ijk}))\phi_{ijk}^{-1} - (r/2)(h(\theta_{ijk}) - \mathbf{x}_{ijk}^T\beta - u_{ij})^2]h'(\theta_{ijk})$.

Once again posterior inference about $g(\theta_{ijk})$'s is performed using (e) and iterated formulas for posterior moments.

To conclude this section, we consider spatial HB GLMs and provide sufficient conditions for the propriety of the posterior. We begin with the likelihood given in (1) and model the θ_{ik} as in (2), but this time the u_i represent variables that if observed would display spatial structure. More particularly, we model the u_i so that a pair of contiguous zones would have stronger (positive) correlation than any arbitrary zones that are noncontiguous.

For $u_1,\ldots,u_m$, we consider the prior

$$p(u_1,\ldots,u_m|r_u) \propto r_u^{m/2}\exp\left[-\frac{r_u}{2}\sum_{i<l} w_{il}(u_i - u_l)^2\right], \tag{6}$$

where the w_{il} are strictly positive if zones i and l are contiguous, and $w_{il} = 0$ otherwise. This prior is a special case of general pairwise difference priors, considered by, for example, Besag, Green, Higdon, and Mengersen (1995). The marginal priors for β, R_u, and R remain as before.

For brevity, write $n_T = \sum_{i=1}^{m} n_i$ and $\bar{\mathbf{x}} = n_T^{-1}\sum_{i=1}^{m}\sum_{k=1}^{n_i}\mathbf{x}_{ik}$. It is assumed that the matrix $\mathbf{X}_0^T = (\mathbf{x}_{11} - \bar{\mathbf{x}},\ldots,\mathbf{x}_{1n_1} - \bar{\mathbf{x}},\ldots,\mathbf{x}_{m1} - \bar{\mathbf{x}},\ldots,\mathbf{x}_{mn_m} - \bar{\mathbf{x}})$ has rank p. We then obtain the following theorem.

Theorem 2. Assume the conditions of Theorem 1, but where now $n_T - p + d > 1$. Then the joint posterior of the θ_{ik} under the spatial prior (6) is proper.

The proof of this theorem is also deferred to the Appendix. For implementing this Bayes procedure via Gibbs sampling, one finds conditional distributions similar to (i)–(v) earlier, with minor modifications to (ii) and (iv).

Remark. It should be noted that if instead of (2), one models the θ_{ik} as

$$h(\theta_{ik}) = \beta_0 + \mathbf{x}_{ik}^T\beta + u_i + \varepsilon_{ik}$$

$$(k = 1,\ldots,n_i; i = 1,\ldots,m),$$

then the posterior of the $h(\theta_{ik})$ fails to be proper. The introduction of the intercept term β_0 creates a nonidentifiability in the posterior, which in turn implies that the joint posterior of the $g(\theta_{ik})$ is also improper.

3. DATA ANALYSIS

3.1 Exposure to Health Hazards Dataset

The analysis of the multicategory dataset mentioned in Section 1, where persons in 15 regions of Canada were asked the question "Have you experienced any negative impact of exposure to health hazards in the workplace," is reported in Table 1 and Figure 1. Here for the kth selected age–sex category within the ith region, p_{ijk} denotes the probability that an individual's response falls in the jth category (where the categories are 1 = yes, 2 = no, 3 = not exposed, and 4 = not applicable or not stated). Within the kth selected age–sex category and the ith region, the Z_{ijk} have a joint multinomial$(t_{ik}; p_{i1k},\ldots,p_{iJk})$ distribution, where $t_{ik} = \sum_j Z_{ijk}$. Fitting model (5) with the Poisson likelihood as described in Section 2, and relabeling k as (a, s) for clarity, the regression equation is

$$\mathbf{x}_{ijk}^T\beta = \mu + \tau_a^A + \tau_s^S + \tau_j^J + \tau_{as}^{AS} + \tau_{aj}^{AJ} + \tau_{sj}^{SJ},$$

where μ is the general effect, τ_a^A is the main effect due to the ath age group, τ_s^S is the main effect due to the sth sex, τ_j^J is the main effect due to the jth category response, τ_{as}^{AS} is the interaction effect of the ath age and sth sex, τ_{aj}^{AJ} is the interaction effect of the ath age and jth category response, and τ_{sj}^{SJ} is the interaction effect sth sex and jth category response. To avoid redundancy, we assume the corner point restrictions

$$\begin{aligned}\tau_1^A &= \tau_1^S = \tau_1^J = \tau_{a1}^{AS} = \tau_{1s}^{AS}\\ &= \tau_{a1}^{AJ} = \tau_{1j}^{AJ} = \tau_{s1}^{SJ} = \tau_{1j}^{SJ} = 0\end{aligned}$$

for all a, s, and j.

Using the extremely vague (but proper) priors for R_u and R determined by setting $a = b = c = d = .002$, we generated 10 parallel Gibbs sampling chains of 2,000 iterations each. Using the 1,000 samples from the latter half of these chains (iterations 1,001–2,000), Table 1 contains the HB estimates, the sample proportions, and the associated standard errors for all four categories in each of the cells cross-classified by $2 \times 2 = 4$ demographic categories for three regions: the smallest, the median, and the largest. Figure 1 shows the sample proportions ("Prop"), traditional logistic regression estimates ("Regr"), and hierarchical Bayes estimates "HB") for all 15 regions for females age 40 or younger. For regions with larger overall sample sizes, shrinkage of the estimates toward the logistic regression estimates within each age–sex category is much smaller than that observed in the smaller regions. For example, Figure 1b shows the HB estimates to be very similar to the logistic regression estimates in the sparsely populated Region 2, whereas Figure 1h shows HB estimates very much like the original sample proportions in populous Region 8. Also, within the kth age–sex category in

Table 1. Impact of Exposure to Health Hazards in the Workplace

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region = 2	*Total n = 294*				
M, Age < 40	Yes	.400	.100	.373	.042
	No	.383	.101	.345	.041
	Not exposed	.150	.119	.199	.031
	NA/NS	.067	.125	.083	.015
F, Age < 40	Yes	.257	.100	.266	.035
	No	.284	.098	.279	.035
	Not exposed	.311	.097	.274	.036
	NA/NS	.148	.107	.181	.026
M, Age ≥ 40	Yes	.111	.111	.184	.028
	No	.153	.109	.176	.027
	Not exposed	.167	.108	.156	.026
	NA/NS	.569	.077	.484	.040
F, Age ≥ 40	Yes	.159	.098	.110	.019
	No	.091	.102	.103	.018
	Not exposed	.125	.010	.134	.022
	NA/NS	.625	.065	.654	.034
Region = 3	*Total n = 740*				
M, Age < 40	Yes	.294	.070	.311	.029
	No	.426	.063	.395	.032
	Not exposed	.203	.075	.186	.023
	NA/NS	.077	.080	.108	.015
F, Age < 40	Yes	.246	.064	.235	.024
	No	.273	.063	.287	.026
	Not exposed	.180	.067	.204	.023
	NA/NS	.301	.062	.274	.026
M, Age ≥ 40	Yes	.156	.069	.154	.019
	No	.150	.069	.165	.020
	Not exposed	.100	.071	.112	.016
	NA/NS	.594	.048	.569	.028
F, Age ≥ 40	Yes	.064	.063	.071	.010
	No	.086	.063	.091	.012
	Not exposed	.111	.062	.099	.013
	NA/NS	.739	.033	.739	.021
Region = 8	*Total n = 1707*				
M, Age < 40	Yes	.274	.047	.279	.021
	No	.360	.044	.362	.023
	Not exposed	.253	.048	.253	.020
	NA/NS	.113	.052	.106	.012
F, Age < 40	Yes	.199	.042	.196	.016
	No	.267	.040	.275	.019
	Not exposed	.289	.040	.295	.019
	NA/NS	.245	.041	.234	.017
M, Age ≥ 40	Yes	.113	.047	.130	.013
	No	.166	.046	.174	.016
	Not exposed	.217	.044	.195	.017
	NA/NS	.504	.035	.501	.022
F, Age ≥ 40	Yes	.087	.042	.076	.009
	No	.123	.041	.110	.011
	Not exposed	.119	.041	.131	.012
	NA/NS	.671	.025	.683	.017

the ith region, the shrinkage is again smaller for categories with larger numbers of responses. For example, of females over age 40 in Region 3, 15/234 (6.41%) answered "yes," compared to 173/234 (73.93%) in the "not applicable/not stated" category. As seen in Table 1, the shrinkage is much larger for the former case, again revealing the adaptive nature of the HB estimates. Finally, note that the standard errors associated with the HB estimates are much smaller than those for the sample proportions.

3.2 Missouri Lung Cancer Dataset

Our second example relates to lung cancer mortality rates in the 115 counties in Missouri during the period 1972–1981. Following the original analysis of this data by Tsutakawa (1985, 1988), we separate the city of St. Louis from the remainder of St. Louis County, which surrounds it. Mortality was classified for each county by sex into four age groups: 45–54, 55–64, 65–74, and 75 and older. The population size for each cell was taken to be the midperiod pop-

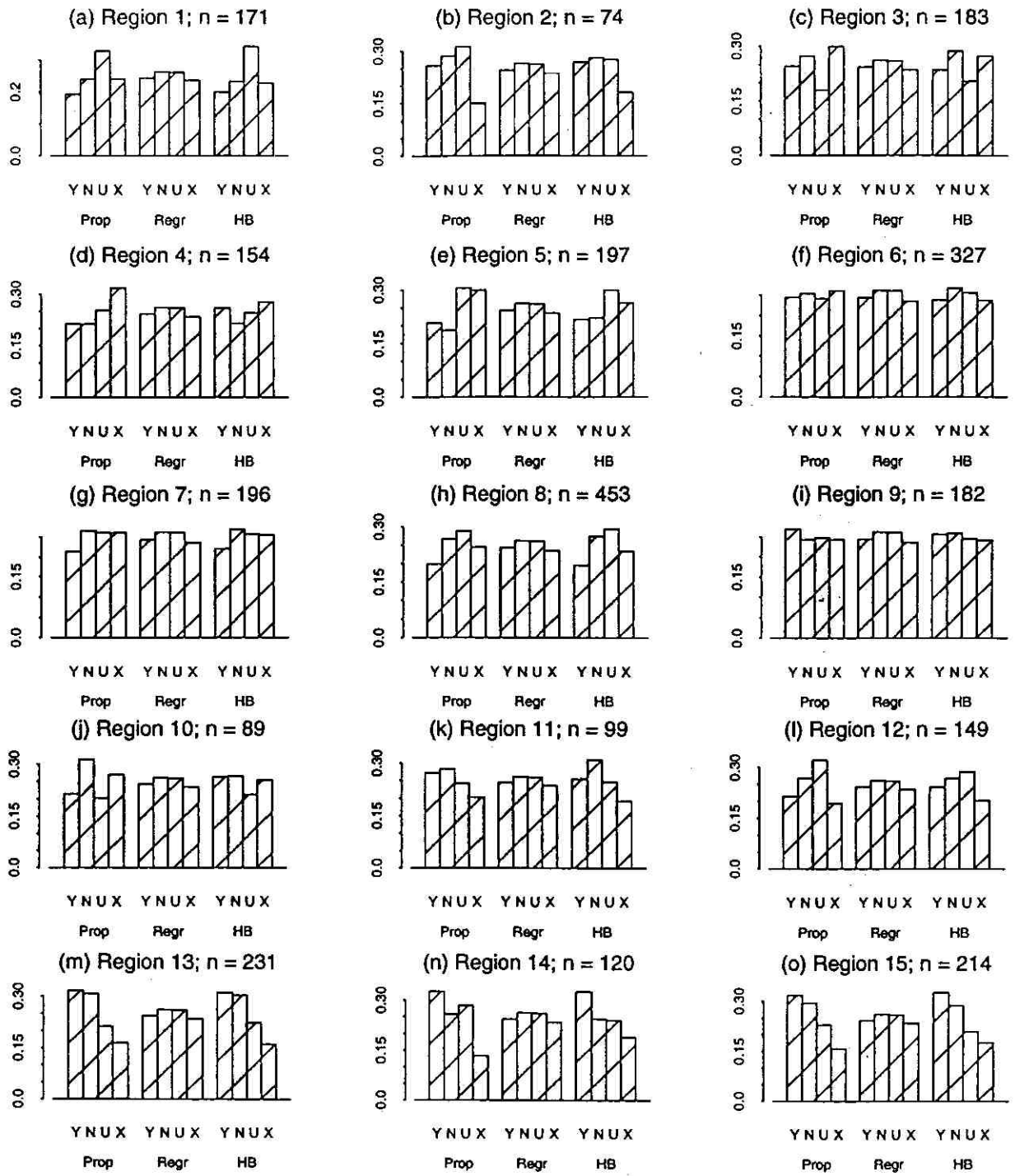

Figure 1. Bar Graph of Estimated Proportions by Category by Region for Females Age < 40. Prop = sample proportion, Regr = logistic regression estimate, HB = hierarchical Bayes estimate. Categories: Y = Yes, N = No, U = Not exposed, X = Not applicable or not stated.

ulation, obtained from the 1970 and 1980 U.S. censuses by linear interpolation.

Again relabeling k as (a, s) for clarity, let Y_{ias} be the lung cancer death count and n_{ias} the midperiod population in the ith county for the ath age group and sth sex, $i = 1, \ldots, 115, a = 1, \ldots, 4, s = 1, 2$. At the first stage of the model, we assume that $Y_{ias}|\zeta_{ias} \overset{\text{ind}}{\sim} \text{Poisson}(\zeta_{ias})$. We then model the mean structure by assuming that $\zeta_{ias} = E_{ias} \exp(\mu_{ias})$, where E_{ias} is the number of deaths that would be expected using some current reference standard and μ_{ias} is the corresponding log-relative risk in cell ias. Some spatial analyses (see, e.g., Bernardinelli and Montomoli 1992) have used an externally available reference table to compute the E_{ias}; here we adopt the simpler alternative of *internal* standardization, defining $E_{ias} = n_{ias} \cdot r$, where $r \equiv \sum_{ias} Y_{ias} / \sum_{ias} n_{ias}$, the statewide lung cancer rate over all sex and age groups in our dataset.

The log-relative risks are then modeled linearly as

$$\mu_{ias} = \mathbf{x}_{as}^T \beta + u_i + \epsilon_{ias}, \tag{7}$$

Table 2. Informal Model Comparison, Missouri Lung Cancer Data

Model for $\mathbf{x}_{as}^T\beta$	Number of fixed effects	Log-likelihood score, $\bar{l}$	Difference
$v_s\alpha + z_a\gamma$	2	580.0	
$v_s\alpha + z_a\gamma + v_s z_a\xi$	3	597.8	17.8
$v_s\alpha + z_a^{(L)}\gamma^{(L)} + v_s z_a^{(L)}\xi^{(L)} + z_a^{(U)}\gamma^{(U)} + v_s z_a^{(U)}\xi^{(U)}$	5	614.9	17.1
$v_s\alpha + z_a^{(L)}\gamma^{(L)} + v_s z_a^{(L)}\xi^{(L)} + z_a^{(M)}\gamma^{(M)} + v_s z_a^{(M)}\xi^{(M)} + z_a^{(U)}\gamma^{(U)} + v_s z_a^{(U)}\xi^{(U)}$	7	618.7	3.8

where β is a vector parameter that captures the effect of sex, age, and sex–age interaction. The ε_{ias} are assumed iid $N(0, \sigma^2)$, but the u_i account for potential spatial clustering of the rates via a conditionally autoregressive (CAR) prior structure (see, e.g., Besag, York, and Mollié 1991; Clayton and Kaldor 1987). That is, we assume that

$$u_i | u_{l\neq i} \sim N(\bar{u}_i, 1/(\tau m_i)),$$

where $\bar{u}_i$ is the average of the $u_{l\neq i}$ that are defined to be "neighbors" of u_i, and m_i is the number of these neighbors. Here we adopt the most common implementation of the CAR structure, defining two counties to be neighbors if and only if they are physically adjacent to each other. It is easy to show that this prior is of the form given in (6), where $w_{il} = 1$ if counties i and l are adjacent, and 0 otherwise. Note that this CAR prior is defined only up to additive constant, again explaining the lack of an intercept term in (7).

It thus remains to determine the appropriate structure for β. Tsutakawa (1988) noted a strong similarity between the male death rates in the two oldest age groups, perhaps due to the competing risks of other diseases. Preliminary analysis of the female rates suggests a similar situation, and as such we begin by defining the sex and age scores

$$v_s = \begin{cases} 0 & \text{if } s = 1 \quad \text{(male)} \\ 1 & \text{if } s = 2 \quad \text{(female)} \end{cases}$$

and

$$z_a = \begin{cases} -1 & \text{if } a = 1 \quad \text{(age 45–54)} \\ 0 & \text{if } a = 2 \quad \text{(age 55–64)} \\ 1 & \text{if } a = 3 \quad \text{(age 65–74)} \\ 1 & \text{if } a = 4 \quad \text{(age 75+)} \end{cases}$$

and use them in a regression-type model,

$$\mathbf{x}_{as}^T\beta = v_s\alpha + z_a\gamma + v_s z_a\xi,$$

thus effectively combining the two oldest age groups. We complete our model specification with flat priors on the components of the fixed effect vector β, a vague gamma(.01, .01) hyperprior on τ, and a moderately informative gamma(1, 1) hyperprior on $R = 1/\sigma^2$. (This latter hyperprior ensures a well-identified joint posterior distribution and, as we shall see, is still quite vague relative to the posterior for the ε_{ias}). We then fit this model via Gibbs sampling using the BUGS language (Spiegelhalter, Thomas, Best, and Gilks 1995), aided by the CODA S+ function (Best, Cowles, and Vines 1995) for assessing convergence and computing posterior summaries. BUGS uses S-like syntax for specifying fairly complex hierarchical models. The program converts this syntax into a directed acyclic graph, the nodes of which correspond to the complete conditional distributions necessary for the Gibbs algorithm. Our results

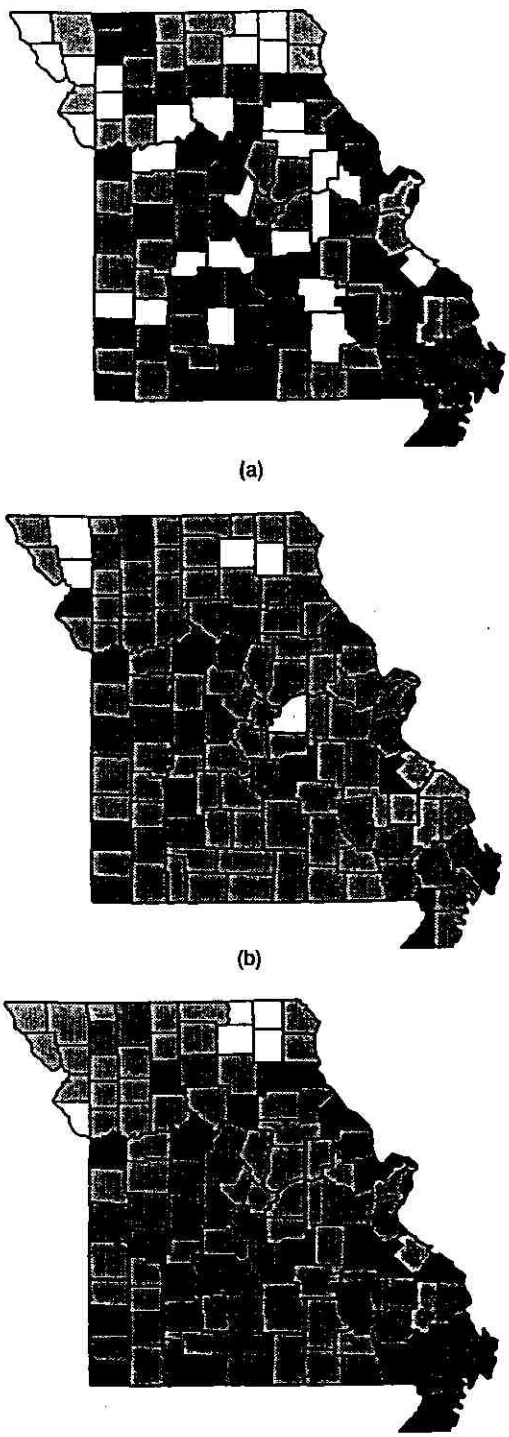

Figure 2. Male 55–64 Lung Cancer Relative Risks, Missouri Counties, 1972–1981. (a) Raw SMRs; (b) Tsutakawa EB smoothed RRs; (c) Spatially smoothed RRs.

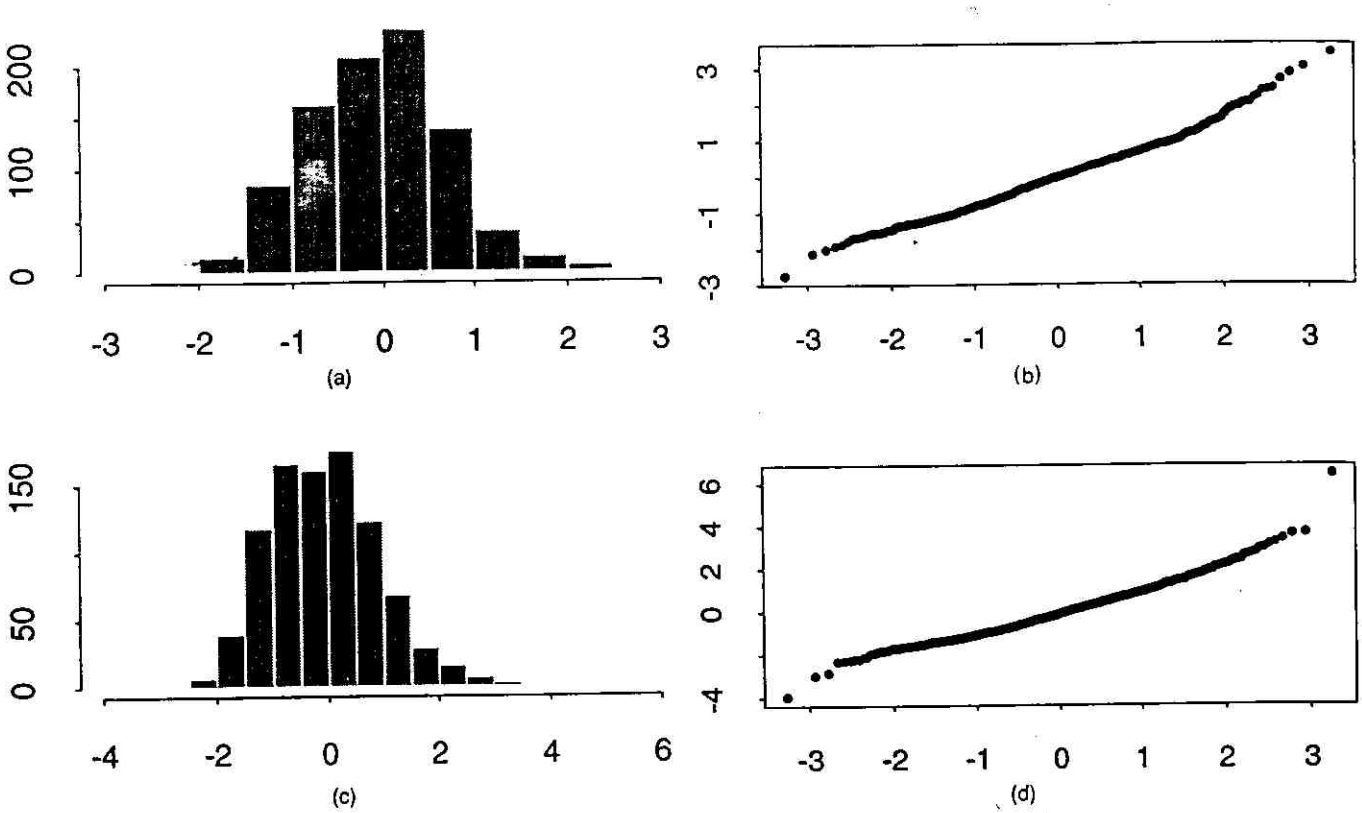

Figure 3. Residual Analysis, Five Fixed-Effects Model, Missouri Lung Cancer Data. (a) Histogram, with overdispersion terms; (b) normal q-q plot, with overdispersion terms; (c) histogram, no overdispersion terms; (d) normal q-q plot, no overdispersion terms.

indicated moderate spatial correlation in the data (posterior for τ centered near 30), a modest need for the extra variability terms (posterior for σ centered near .2), and significant sex–age interaction (posterior for ξ removed from 0).

To investigate the scope of models that our data could support, we considered a simpler model that drops the interaction term ξ and a more complex analysis of variance (ANOVA)-type model that replaced the age score vector $\mathbf{z} = (-1, 0, 1, 1)'$ with separate vectors for the lowest and highest age groups, namely $\mathbf{z}^{(L)} = (1, 0, 0, 0)'$ and $\mathbf{z}^{(U)} = (0, 0, 1, 1)'$. Table 2 compares the fit of these models using the posterior log-likelihood score, computed as the sample average $\bar{l} = 1/G \sum_{g=1}^{G} l^{(g)}$, where

$$l^{(g)} = \sum_{ias} \mu_{ias}^{(g)} y_{ias} - \sum_{ias} E_{ias} \exp(\mu_{ias}^{(g)}) + C,$$

$$g = 1, \ldots, G.$$

Here the superscript (g) indexes the Gibbs iterates, and C is a scaling constant. After a burn-in period of 50 iterations, we found that retaining $G = 500$ iterations was sufficient to produce log-likelihood scores with batched standard errors near .5. Note that the average score $\bar{l}$ for the model with five fixed effects is larger than that for the model with three, which in turn is substantially larger than that for the two fixed-effects model. However, a final extension to the saturated model that separates the two oldest age groups—that is, using $\mathbf{z}^{(M)} = (0, 0, 1, 0)'$ and $\mathbf{z}^{(U)} = (0, 0, 0, 1)'$—offers no numerically significant improvement in fit. Although the usual chi-squared asymptotics for differences in $-2\bar{l}$ are not appropriate in our Bayesian random-effects model setting, it seems clear from Table 2 that the model with five fixed effects offers the best fit while preserving parsimony.

Our chosen model produces posterior means and 95% equal-tail credible sets as follows: for $\alpha, -1.46$ and $(-1.545, -1.36)$; for $\gamma^{(L)}, -1.064$ and $(-1.15, -.976)$; for $\gamma^{(U)}$, .558 and (.503, .630); for $\xi^{(L)}$, .369 and (.227, .503); and for $\xi^{(U)}, -.318$ and $(-.428, -.207)$. Thus log-relative risk is nearly 1.5 units lower for females than for males on average, with the risk increasing monotonically with age. (Recall that the two oldest age groups have been combined.) However, the signs on the interaction terms $\xi^{(L)}$ and $\xi^{(U)}$ show that this increase is not as dramatic for females as for males.

Figure 2 maps the raw standardized mortality ratios for men age 55–64, $\text{SMR}_{i21} = Y_{i21}/E_{i21}$, the fitted relative risks obtained by Tsutakawa (1988) using EB methods without a spatial smoothing prior, and the fitted relative risks from our fully Bayesian spatial smoothing analysis, $\text{RR}_{i21} = 1/G \sum_{g=1}^{G} \exp(u_i^{(g)} + \varepsilon_{i21}^{(g)})$, the average of the $G = 500$ corresponding postconvergence relative risk estimates. Although the comparison between our results and Tsutakawa's is not completely fair, because the latter were obtained using data for males only, clearly both of these methods eliminate much of the noise in the original map while preserving the high rate in populous St. Louis city. However, our spatial model clarifies the general increase in rates from north to south (especially along the eastern border with Illinois) and also identifies possible clusters of counties with similar risk, while maintaining a reasonable amount of fidelity to the original data.

Finally, we check our model by analyzing the posterior means of the collection of standardized residuals, $r_{ias} = E[(Y_{ias} - \zeta_{ias})/\sqrt{\zeta_{ias}}|\mathbf{y}]$, which are readily computable in BUGS (Spiegelhalter et al. 1995, pp. 40–46). Figure 3a shows a histogram of these mean residuals, and Figure 3b gives their normal Q-Q plot. Both plots reveal a high degree of normality.

Finally, the rather small fitted standard deviation (.17) for the extra viability terms ε_{ias} made us wonder whether these terms were even needed in the model. To check this, we reran our model without these terms, obtaining the residual histogram and normal plot shown in Figures 3c and 3d. Although the degree of normality is still acceptable, the presence of a few large outliers is disturbing. The one enormous outlier on the high side corresponds to men in the youngest age group who live in the city of St. Louis; apparently their very high lung cancer death rate is poorly fit by the model. Interestingly, two of the three outlying values on the low side are the youngest and second-youngest groups of men living in St. Louis *County*, who are apparently much healthier than the model predicts. Thus we conclude that the overdispersion terms ε_{ias} are critical in obtaining acceptable fits in all-urban St. Louis city and its only geographic neighbor, suburban St. Louis County, allowing differing rates in these two disparate regions despite their juxtaposition on the map.

4. CONCLUSIONS

In this article we have provided a general approach for small-area estimation based on hierarchical Bayes generalized linear models, with and without spatial correlation structure. Sufficient conditions have been given to ensure the propriety of posteriors under noninformative priors. The general methodology is applicable to a wide variety of situations calling for simultaneous estimation of small-area parameters. Future work looks to continued automation in the fitting of these models via MCMC methods, especially in the areas of model choice and model averaging. Promising tools in this regard include expected predicted deviance scores, recently introduced by Gelfand and Ghosh (1997) and illustrated for spatio-temporal models by Waller, Carlin, Xia, and Gelfand (1997).

APPENDIX: PROOFS

Proof of Theorem 1

The joint posterior pdf of $\theta, \beta, \mathbf{u}, R_u$, and R given $\mathbf{y}$ is

$$\begin{aligned}
&\pi(\theta, \beta, \mathbf{u}, r_u, r|\mathbf{y}) \\
&\propto \prod_i \prod_k \exp[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))] r^{1/2 \sum_1^m n_i} \\
&\times \prod_i \prod_k \exp\left[-\frac{r}{2}(h(\theta_{ik}) - \mathbf{x}_{ik}^T\beta - u_i)^2\right] \\
&\times \left(\prod_i \prod_k h'(\theta_{ik})\right) r_u^{m/2} \exp\left(-\frac{r_u}{2}\sum_1^m u_i^2\right) \\
&\times \exp\left(-\frac{ar_u}{2}\right) r_u^{1/2b-1} \exp\left(-\frac{cr}{2}\right) r^{1/2d-1}.
\end{aligned}$$

Integrating with respect to β, r_u, and r in succession, we obtain

$$\begin{aligned}
\pi(\theta, \mathbf{u}|\mathbf{y}) &\le C \prod_i \prod_k \exp[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))] \\
&\times \left(a + \sum_1^m u_i^2\right)^{-1/2(m+b)} \prod_i \prod_k h'(\theta_{ik}),
\end{aligned}$$

where C (> 0) is a generic constant that does not depend on θ or $\mathbf{u}$. Now integrating with respect to $\mathbf{u}$ and using the structure of a multivariate t, it follows that

$$\pi(\theta|\mathbf{y}) \le C \prod_i \prod_k \exp[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))] h'(\theta_{ik}).$$

The result now follows from (3).

Proof of Theorem 2

For notational simplicity, without loss of generality h is taken as the identity function throughout. The joint posterior of $\theta, \beta, \mathbf{u}, R_u$, and R given $\mathbf{y}$ is

$$\begin{aligned}
&\pi(\theta, \beta, \mathbf{u}, r_u, r|\mathbf{y}) \\
&\propto \prod_i \prod_k \exp[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))] r^{n_T/2} \\
&\times \prod_i \prod_k \exp\left[-\frac{r}{2}(\theta_{ik} - \mathbf{x}_{ik}^T\beta - u_i)^2\right] \\
&\times r_u^{m/2} \exp\left[-\frac{r_u}{2}\sum_{1 \le i < l \le m} w_{il}(u_i - u_l)^2\right] \\
&\times \exp\left(-\frac{ar_u}{2}\right) r_u^{1/2b-1} \exp\left(-\frac{cr}{2}\right) r^{1/2d-1}.
\end{aligned}$$

With the one-to-one transformation $(z_1, \ldots, z_{m-1}, u_m)$, where $z_i = u_i - u_m, i = 1, \ldots, m$, the posterior transforms to

$$\begin{aligned}
&\pi(\theta, \beta, u_m, \mathbf{z}, r_u, r|\mathbf{y}) \\
&\propto \exp\left[\sum_i \sum_k \phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))\right] r^{n_T/2} \\
&\times \exp\left[-\frac{r}{2}\sum_i \sum_k (\theta_{ik} - \mathbf{x}_{ik}^T\beta - z_i - u_m)^2\right] \\
&\times \exp\left[-\frac{r_u}{2}\left\{a + \sum_{1 \le i < l \le m} w_{il}(z_i - z_l)^2\right\}\right] r_u^{(m+b-1)/2} \\
&\times \exp\left(-\frac{cr}{2}\right) r^{1/2d-1},
\end{aligned}$$

where $z_m = 0$ and $\mathbf{z} = (z_1, \ldots, z_{m-1})$. Next, write $\bar{\theta} = n_T^{-1}\sum_i \sum_k \theta_{ik}$ and $\bar{z} = m^{-1}\sum_i z_i$. Integrating with respect to u_m, β, r_u, and r in succession, we have

$$\begin{aligned}
\pi(\theta, \mathbf{z}|\mathbf{y}) &\le C \exp\left[\sum_i \sum_k \phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))\right] \\
&\times \left[a + \sum_{1 \le i < l \le m} w_{il}(z_i - z_l)^2\right]^{-(m+b)/2},
\end{aligned}$$

where C (> 0) is a generic constant that does not depend on θ or $\mathbf{z}$. Recall that $z_m = 0$ and $\sum_{1 \le i < l \le m} w_{il}(z_i - z_l)^2$ involves only m

$-$ 1 variables $z_1, \ldots, z_{m-1}$. Thus, integrating with respect to z, and using the structure of a multivariate t distribution yields

$$\pi(\theta|y) \leq C \exp\left[\sum_i \sum_k \phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik}))\right].$$

The result again follows from (3).

[Received December 1995. Revised June 1997.]

REFERENCES

Albert, J. H. (1988), "Computational Methods Using a Bayesian Hierarchical Generalized Linear Model," *Journal of the American Statistical Association*, 83, 1037–1044.

Bernardinelli, L., and Montomoli, C. (1992), "Empirical Bayes Versus Fully Bayesian Analysis of Geographical Variation in Disease Risk," *Statistics in Medicine*, 11, 983–1007.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Besag, J., York, J. C., and Mollié, A. (1991), "Bayesian Image Restoration, With Two Applications in Spatial Statistics" (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Best, N. G., Cowles, M. K., and Vines, K. (1995), "CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30," technical report, Cambridge University, Institute of Public Health, Medical Research Council Biostatistics Unit.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Brooks, R. J. (1984), "Approximate Likelihood Ratio Tests in the Analysis of Beta-Binomial Data," *Applied Statistics*, 33, 285–289.

Clayton, D. G., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–681.

Datta, G. S., and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *The Annals of Statistics*, 19, 1748–1770.

Dellaportas, P., and Smith, A. F. M. (1993), "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling," *Applied Statistics*, 42, 443–459.

Dempster, A. P., and Tomberlin, T. J. (1980), "The Analysis of Census Undercount From a Post-Enumeration Survey," in *Proceedings of the Conference on Census Undercount*, pp. 88–94.

Farrell, P., MacGibbon, B., and Tomberlin, T. J. (in press), "Empirical Bayes Estimators of Small Area Proportions in Multistage Designs," submitted to *Statistica Sinica*.

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

Gelfand, A. E., and Ghosh, S. K. (in press), "Model Choice: A Minimum Posterior Predictive Loss Approach," submitted to *Biometrika*.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–511.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Ghosh, M., and Lahiri, P. (1987), "Robust Empirical Bayes Estimation of Means From Stratified Samples," *Journal of the American Statistical Association*, 82, 1153–1162.

——— (1992), "A Hierarchical Bayes Approach to Small Area Estimation With Auxiliary Information," in *Bayesian Analysis in Statistics and Econometrics*, Lecture Notes in Statistics 75, eds. P. K. Goel and N. S. Iyengar, New York: Springer-Verlag, pp. 107–125.

Ghosh, M., and Meeden, G. (1986), "Empirical Bayes Estimation in Finite Population Sampling," *Journal of the American Statistical Association*, 74, 269–277.

Ghosh, M., and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal" (with discussion), *Statistical Science*, 9, 65–93.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Leonard, T., and Novick, M. R. (1986), "Bayesian Full Rank Marginalization for Two-Way Contingency Tables," *Journal of Educational Statistics*, 11, 33–56.

MacGibbon, B., and Tomberlin, T. J. (1989), "Small Area Estimates of Proportions via Empirical Bayes Techniques," *Survey Methodology*, 15, 237–252.

Maiti, T. (1997), "Hierarchical Bayes Estimation of Mortality Rates for Disease Mapping," Technical Report 546, University of Florida, Dept. of Statistics.

Malec, D., Sedransk, J., and Tompkins, L. (1993), "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interview Survey," in *Case Studies in Bayesian Statistics*, eds. C. Gatsonis, J. S. Hodges, R. E. Kass, and N. D. Singpurwalla, New York: Springer-Verlag, pp. 377–389.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Nandram, B., and Sedransk, J. (1993), "Bayesian Predictive Inference for a Finite Population Proportion: Two-Stage Cluster Sampling," *Journal of the Royal Statistical Society*, Ser. B, 55, 399–408.

Prasad, N. G. N., and Rao, J. N. K. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163–171.

Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. R. (1995), "BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50," technical report, Cambridge University, Institute of Public Health, Medical Research Council Biostatistics Unit.

Statistics Canada (1992), "The 1991 General Social Survey—Cycle 6: Health," in *Public Use Microdata File Documentation and User's Guide*, Ottawa: Statistics Canada.

Stroud, T. W. F. (1987), "Bayes and Empirical Bayes Approaches to Small Area Estimation of Small Area Statistics," in *International Symposium on Small Area Statistics*, eds. R. Platek, J. N. K. Rao, C. E. Saerndal, and M. P. Singh, New York: Wiley, pp. 124–140.

——— (1991), "Hierarchical Bayes Predictive Means and Variances With Application to Sample Survey Inference," *Communications in Statistics, Part A—Theory and Methods*, 20, 13–36.

——— (1994), "Bayesian Inference From Categorical Survey Data," *Canadian Journal of Statistics*, 22, 33–45.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tsutakawa, R. K. (1985), "Estimation of Cancer Mortality Rates: A Bayesian Analysis of Small Frequencies," *Biometrics*, 41, 69–79.

——— (1988), "Mixed Model for Analyzing Geographic Variability in Mortality Rates," *Journal of the American Statistical Association*, 83, 37–42.

Tsutakawa, R. K., Shoop, G. L., and Marienfeld, C. J. (1985), "Empirical Bayes Estimation of Cancer Mortality Rates," *Statistics in Medicine*, 4, 201–212.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997), "Hierarchical Spatio-Temporal Mapping of Disease Rates," *Journal of the American Statistical Association*, 92, 607–617.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach

Malay GHOSH, Narinder NANGIA, and Dal Ho KIM

This article develops a general methodology for small domain estimation based on data from repeated surveys. The results are directly applied to the estimation of median income of four-person families for the 50 states and the District of Columbia. These estimates are needed by the U.S. Department of Health and Human Services (HHS) to formulate its energy assistance program for low income families. The U.S. Bureau of the Census, by an informal agreement, has provided such estimates to HHS through a linear regression methodology since the latter part of the 1970s. The current method is an empirical Bayes method (EB) that uses the Current Population Survey (CPS) estimates as well as the most recent decennial census estimates updated by the per capita income estimates of the Bureau of Economic Analysis. However, with the existing methodology, standard errors associated with these estimates are not easy to obtain. The EB estimates, when used naively, can lead to underestimation of standard errors. Moreover, because the sample estimates are collected through the CPS every year, there is a very natural time series aspect of the data that is currently ignored. We have performed a full Bayesian analysis using a hierarchical Bayes (HB) time series model. In addition to providing the median income estimates as the posterior means, we have provided also the posterior standard deviations. Included in our model is the information on the median incomes of three- and five-person families as well. In this way a multivariate HB procedure is used. The Bayesian analysis requires evaluation of high-dimensional integrals. We have overcome this problem by using the Gibbs sampling technique, which has turned out to be a very convenient tool for Monte Carlo integration. Also, we have validated our results by comparing them against the 1989 four-person median income figures obtained from the 1990 census. We used four different criteria for such comparisons. It turns out that the estimates obtained by using a bivariate time-series model are the best overall. We use a criterion based on deviances for model selection and also provide a sensitivity analysis of the proposed hierarchical model.

KEY WORDS: Current Population Survey; Empirical Bayes; First-order autoregressive; Hierarchical Bayes; Multivariate; Small area estimation.

1. INTRODUCTION

Estimates of median incomes of four-person families at the national, state, county, and local area levels are often needed for a variety of governmental decisions. The U.S. Department of Health and Human Services (HHS) has a direct need for such data at the state level (the 50 states and the District of Columbia) for formulating its energy assistance program to low income families. Such estimates are provided to the HHS annually by the Bureau of the Census.

First, we discuss briefly the current approach of the Bureau of the Census for producing such estimates. (The details appear in Fay, Nelson, and Litow 1993). This methodology relies on three sources of data. The basic source is the annual demographic supplement to the March sample of the Current Population Survey (CPS), which provides annually median income by states for families of different sizes. Second, once every 10 years, similar figures are obtained from the decennial census for the year proceeding the census year; for example 1969, 1979, 1989, and so on. Third, the Bureau of the Census uses also annual estimates of the per capita income (PCI) obtained by the Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce.

Direct use of the CPS estimates is limited due to the smallness of the sample size, which causes substantial variability. In contrast, the census estimates based on the long-form sample (filled in by approximately $\frac{1}{6}$ of the population) are believed to have virtually negligible standard errors and can be used in conjunction with the CPS estimates in producing the annual median income estimates. The census estimates are also used as the "gold standard" against which other estimates are tested. Such a comparison, however, is only possible for those years that immediately precede the census year. Finally, the PCI estimates, unlike the CPS estimates, do not have associated sampling errors, as they are not obtained using sampling techniques.

The current Bureau of the Census approach uses a bivariate regression model as suggested by Fay (1987). This method includes median incomes of three- and five-person families along with those for four-person families, although the primary objective continues to be estimation of median incomes of four-person families. For each state, based on the direct CPS estimates, the median incomes for three-, four-, or five-person families are obtained by linear interpolation using tabulated income categorized into intervals of $2,500. The basic data set for each state is a bivariate random vector with one component equal to the median income of four-person families and the other component equal to a weighted average of median incomes of three- and five-person families, the weights being .75 and .25.

This research was partially supported by National Science Foundation grants SES-9201210 and SBR-9423996 and a Joint Statistical Agreement with the Bureau of the Census. The views expressed herein reflect those of the authors and not of the Bureau of the Census. Malay Ghosh is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Narinder Nangia is Statistical Consultant, Trilogy Corporation, Waukeegan, IL 60085. Dal Ho Kim is Instructor, Department of Statistics, Kyungpook National University, Taegu 702-701, Korea. The authors gratefully acknowledge Robert Fay of the Bureau of the Census for introducing them to this problem and for many helpful discussions throughout the course of this investigation. The article has benefitted much from the constructive suggestions of William Bell of the Bureau of the Census, the associate editor, and three referees. Special thanks to Larry Winner for his help in the preparation of the figures.

Journal of the American Statistical Association
December 1996, Vol. 91, No. 436, Applications and Case Studies

In addition to the intercept term, the regression equation used by the Bureau of the Census uses as independent variables the base year census median (b) and the adjusted census median (c) both for four-person families, as well as the weighted average of three- and five-person families with the same .75 and .25 weights. Here census median (b) is a generic symbol for the median income of a family of particular size in a state from the recently most available decennial census. The adjusted census median (c) is obtained from the following formula:

$$\text{Adjusted census median } (c) = [\text{BEA PCI}(c)/\text{BEA PCI}(b)] \times \text{census median } (b), \quad (1)$$

where BEA PCI(c) and BEA PCI(b) represent PCI estimates produced by the BEA for the current year c and the base-year b. As pointed out by Fay (1987), formula (1) attempts to adjust the base year census median by the proportional growth in the BEA PCI to arrive at the current year adjusted median. The inclusion of the census median (b) as a second independent variable is believed to adjust for any possible overstatement of the effect of change in BEA PCI in estimating the current median incomes.

Finally, weighted averages of the CPS sample estimate of the current median income and the corresponding regression estimates are obtained. These weighted estimates are obtained by using an empirical Bayes (EB) procedure (Fay 1987; Fay et al. 1993) with a somewhat ad hoc estimator of the prior variance.

In an earlier paper (Datta, Ghosh, Nangia, and Natarajan 1996), the ideas of Fay (1987) were modified, extended, and implemented. A more appealing EB procedure was given estimating this variance by its maximum likelihood estimator (MLE), based on the marginal distributions of the observations. Second, in addition to this EB procedure, full Bayesian solutions were offered for the same estimation problem using both univariate and multivariate hierarchical Bayes (HB) models. Although the univariate procedure utilized only the median income of four-person families, the different multivariate procedures also utilized the median incomes of three- and five-person families in various ways.

A comparison of the estimates and the corresponding census figures for the income year 1979 revealed that both the HB and the EB procedures improved tremendously over the CPS medians under both the univariate and the multivariate models. We also found that the point estimates obtained by using either the univariate model or some version of a multivariate model did not substantially differ. However, the standard errors and the coefficients of variation were reduced considerably by using a multivariate model in comparison with the univariate model. Also, we observed that the EB procedure resulted in underestimation of standard errors in contrast to a HB procedure. The familiar explanation of this phenomenon is that an EB procedure based on estimated priors fails to account for the uncertainty involved in the estimation of prior parameters and can often lead to underestimates of standard errors. A HB method accounts for this uncertainty by assigning distributions (albeit often diffuse ones) to the prior parameters.

The methodology suggested in this article goes yet one step further. Because of the repetitive nature of the CPS, it seems possible to obtain better estimates of statewide medians by Bayesian time series modeling. We demonstrate this by finding estimates of statewide median incomes of four-person families for 1989, using 1979 as the base year. We compare both the time series and non–time series estimates with the CPS estimates as well as with the EB estimates of the Bureau of the Census in the light of the decennial census figures.

Several HB models, both time series and non–time series, were tried. Half of these used normal regression models, and the other half used lognormal regression models. The normal models outperformed the lognormal models in all circumstances.

For each normal time series or non–time series case, we considered three separate regression models where the intercept term was always included. Among the three cases, one included only the adjusted census median (c) as the independent variable, the second included only the base year census median (b) as the independent variable, and the third included both the adjusted census median and the base year census median as independent variables.

We compared all of these estimates to the 1989 decennial census estimates. The results turned out to be quite interesting, especially in view of what was presented by Fay (1987). First, it turned out that under all circumstances, a regression model utilizing only adjusted census medians as covariates was performing better than those including either the base year census median as a covariate or both the base year and adjusted census medians as covariates. This is in contrast to Fay's (1987) recommendation to include both the adjusted and base year census medians as covariates, as was evidenced from the 1979 figures. Second, it turned out that the bivariate time series model that included the median incomes of four- and five-person families performed the best. This is different from the bivariate model of Fay, which included the weighted average of median incomes of three- and five-person families as the second variable. Finally, though not very surprising, we found that the time series models always outperformed their non–time series counterparts.

A simple model selection based on deviances (defined Sec. 3) along the lines of Malec and Sedransk (1994) lent further support to the approach of retaining only the adjusted census medians as covariates along with the intercept term. It turned out that the corresponding deviance term was quite close to the one based on the saturated model that included both the adjusted and the nonadjusted census medians as covariates along with the intercept term. On the other hand, the model that used only the nonadjusted census median as a covariate along with the intercept term had a deviance quite different from the one resulting from the saturated model.

The outline of the remaining sections is as follows. Section 2 introduces a HB multivariate time series model that can be used not only for the problem at hand, but also for other similar problems. Gibbs sampling is used for computing the necessary estimates and the standard errors for

the parameters of interest. Some general remarks are made about the implementation of the Gibbs sampler. The posterior distributions needed for implementing the Gibbs sampler are given in the Appendix.

Section 3 addresses the specific problem of estimation of median income of four-person families using univariate, bivariate, and trivariate time series and non–time series models. Our findings are summarized in three graphs and several tables. We give the details for the case in which the adjusted census medians are used as the only auxiliary variables along with the intercept terms, and compare these estimates to the figures obtained from the 1990 census according to four criteria introduced in Section 3. As discussed in the preceding paragraph, our findings indicate that a bivariate time series model in which one component variable is the median income of five-person families and the other is the median income of four-person families performs best under each one of these four criteria. Section 3 also contains some model selection. Section 4 presents a sensitivity analysis related to departure from the proposed hierarchical model. Finally, Section 5 provides some concluding remarks.

Before ending this section, we reemphasize that the HB method enables us to report not only point estimates, but also the standard errors associated with these estimates. This is a distinct advantage of the HB methodology over the EB methodology currently used.

2. THE GENERAL MULTIVARIATE HIERARCHICAL BAYES MODEL

Suppose, based on a given sample at time j ($j = 1, \ldots, t$), that $\mathbf{Y_{ij}} = (Y_{ij1}, \ldots, Y_{ijs})^T$ is a s-dimensional column vector of sample survey estimators of some characteristics $\boldsymbol{\theta}_{\mathbf{ij}} = (\theta_{ij1}, \ldots, \theta_{ijs})^T$, for the ith small area ($i = 1, \ldots, m$). The problem is to estimate some function of the $\boldsymbol{\theta}_{\mathbf{ij}}$'s. In the specific problem of estimation of median income of four-person families at time u, writing θ_{ij1} as the median income of four-person families at time j for the ith local area, we are interested in estimating $(\theta_{1u1}, \cdots, \theta_{mu1})^T$. Clearly, there could be other parameters of interest. For example, we may be interested in estimating $(\theta_{1v1} - \theta_{1u1}, \ldots, \theta_{mv1} - \theta_{mu1})^T$, the change in the median income of four-person families from time u to time v.

The general HB model is as follows:

I. $\mathbf{Y}_{ij}|\boldsymbol{\theta}_{ij} \overset{\text{ind}}{\sim} N(\boldsymbol{\theta}_{ij}, \mathbf{V}_{ij})$ $(i = 1, \ldots, m; j = 1, \ldots, t)$, where $\mathbf{V}_{ij}$'s are known

II. $\boldsymbol{\theta}_{ij}|\boldsymbol{\alpha}, \mathbf{b}_j, \boldsymbol{\psi}_j \overset{\text{ind}}{\sim} N(\mathbf{X}_{ij}\boldsymbol{\alpha} + \mathbf{Z}_{ij}\mathbf{b}_j, \boldsymbol{\psi}_j)$ $(i = 1, \ldots, m; j = 1, \ldots, t)$

III. $\mathbf{b}_j|\mathbf{b}_{j-1}, \mathbf{W} \overset{\text{ind}}{\sim} N(\mathbf{b}_{j-1}, \mathbf{W})$ $(j = 1, \ldots, t)$

IV. Marginally $\boldsymbol{\alpha}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_t$, and $\mathbf{W}$ are mutually independent with

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{uniform } (R^p), \\ \boldsymbol{\psi}_j &\sim \text{inverse Wishart } (\mathbf{S}_j, k_j), \quad \text{and} \\ \mathbf{W} &\sim \text{inverse Wishart } (\mathbf{S}_0, k_0). \end{aligned}$$

In II, $\mathbf{X}_{ij}(s \times p)$ and $\mathbf{Z}_{ij}(s \times q)$ are known design matrices.

For the particular problem at hand, the assumption of conditional independence of the $\mathbf{Y}_{ij}$ given the $\boldsymbol{\theta}_{ij}$ may be open to question. But, as we see in Section 4, the independence model works better than some of the competing first-order autoregressive models.

It is possible to allow diffuse priors for $\boldsymbol{\alpha}, \boldsymbol{\psi}_j$, and $\mathbf{W}$ as long as the posterior distribution of $\boldsymbol{\theta}_{ij}$ given $\mathbf{Y}_{ij}$ $(i = 1, \ldots, m, j = 1, \ldots, t)$ remains proper. It may be tempting to combine stages I and II of the model and write

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\alpha} + \mathbf{Z}_{ij}\mathbf{b}_j + \mathbf{u}_{ij} + \mathbf{e}_{ij},$$

where the $\mathbf{u}_{ij}$'s and $\mathbf{e}_{ij}$'s are mutually independent with $\mathbf{u}_{ij} \overset{\text{ind}}{\sim} N(0, \boldsymbol{\psi}_j)$ and $\mathbf{e}_{ij} \overset{\text{ind}}{\sim} N(0, \mathbf{V}_{ij})$. This is clearly a mixed-effect multi-variate analysis of variance (MANOVA) model. But this rewriting, though helpful for inference about $\boldsymbol{\alpha}$ and $\mathbf{b}_j$, does not help directly for inference about $\boldsymbol{\theta}_{ij}$.

Part II of our model bears a strong similarity to the observational equations in a dynamic linear model (see, e.g., Broemeling 1985), although the $\boldsymbol{\theta}_{ij}$'s themselves are not observables. Part III of the model corresponds the systems equations in dynamic linear models. An alternative way of writing this is $\mathbf{b}_j = \mathbf{b}_{j-1} + \mathbf{z}_j$, where $\mathbf{z}_j$ are iid $N(0, \mathbf{W})$. This is the so-called random walk model, which has been used quite extensively by time series analysts (see, e.g., Bell 1984). If the variance matrices $\boldsymbol{\psi}_j$'s and $\mathbf{W}$ were known, then standard Bayesian analysis for dynamic linear models could be performed using the Kalman filter updating algorithm (see, e.g., Meinhold and Singpurwalla 1983 and West, Harrison, and Migon 1985). However, in the absence of knowledge of the $\boldsymbol{\psi}_j$'s and $\mathbf{W}$, the advantage of using a Kalman filter is lost, and direct Bayesian analysis must be performed. The objective of this analysis is to find the posterior distributions of the $\boldsymbol{\theta}_{ij}$'s given the data $\mathbf{y}_{ij}$ $(i = 1, \ldots, m; j = 1, \ldots, t)$. Such distributions are analytically intractable and require high-dimensional numerical integration. Instead, we adopt Monte Carlo integration and use Gibbs sampling.

Gibbs sampling, originally introduced by Geman and Geman (1984) and more recently popularized by Gelfand and Smith (1990), is a Markovian updating scheme that requires sampling from full conditional distributions. Densities (which could be multivariate) are denoted generally by square brackets so that the joint, conditional, and marginal densities appear as, for example, $[U, V]$, $[U|V]$, and $[V]$. Given a collection of random variables (real or vector valued) $U_1, \ldots, U_k$, the joint density $[U_1, \ldots U_k]$ is assumed to be uniquely determined by $[U_s|U_r, r \neq s], s = 1, \ldots k$. The interest is in finding the marginal distributions $[U_s], s = 1, \ldots, k$.

For the model given in I–IV, the full conditionals determine the joint pdf of $\boldsymbol{\theta}_{ij}$ given $\mathbf{y}_{ij}$ uniquely. In implementing the Gibbs sampler, we follow the recommendation of Gelman and Rubin (1992) and run $n(\geq 2)$ parallel chains, each for $2d$ iterations with starting points drawn from an overdispersed distribution. But to diminish the effects of the starting distributions, the first d iterations of each chain are discarded. After d iterations, all the subsequent iterates are retained for finding the desired posterior distributions, pos-

terior means, and variances, as well as for monitoring the convergence of the Gibbs sampler. The convergence monitoring is discussed in greater detail in Section 3.

To implement the Gibbs sampler, we need to generate samples from the full conditional distributions of

$$\alpha|\mathbf{y}, \theta, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W};$$

$$\mathbf{b}_j|\mathbf{y}, \theta, \alpha, \mathbf{b}_1, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_{j+1}, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W} \quad (2 \le j \le t-1);$$

$$\mathbf{b}_1|\mathbf{y}, \theta, \alpha, \mathbf{b}_2, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W};$$

$$\mathbf{b}_t|\mathbf{y}, \theta, \alpha, \mathbf{b}_1, \ldots, \mathbf{b}_{t-1}, \psi_1, \ldots, \psi_t, \mathbf{W};$$

$$\psi_j|\mathbf{y}, \theta, \mathbf{b}_1, \ldots, \mathbf{b}_t, \mathbf{W}, \psi_1, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_t;$$

$$\mathbf{W}|\mathbf{y}, \theta, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t;$$

and

$$\theta_{ij}|\mathbf{y}, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W}, \theta_{kl}((k,l) \ne (i,j)),$$

where $\mathbf{y} = (\mathbf{y}_{11}^T, \ldots, \mathbf{y}_{mt}^T)^T$. These distributions are given in the Appendix.

Using the Gibbs sampler, the posterior distribution of θ_{ij} given $\mathbf{y}$ is approximated by

$$\Pi(\theta_{ij}|\mathbf{y}) \approx (nd)^{-1} \sum_{k=1}^{n} \sum_{l=d+1}^{2d} [\theta_{ij}|\mathbf{y}, \alpha = \alpha_{kl}, \mathbf{W} = \mathbf{W}_{kl}, \mathbf{b}_j = \mathbf{b}_{jkl}, \psi_j = \psi_{jkl}, j = 1, \ldots, t]. \quad (2)$$

Also, following Gelfand and Smith (1991), "Rao–Blackwellized" estimates of posterior means and variances of the θ_{ij} are given by

$$E(\theta_{ij}|\mathbf{y} \approx (nd)^{-1} \sum_{k=1}^{n} \sum_{l=d-1}^{2d} (\mathbf{V}_{ij}^{-1} + \psi_{jkl}^{-1})^{-1} \times [\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_{jkl}^{-1}(\mathbf{X}_{ij}\alpha_{kl} + \mathbf{Z}_{ij}\mathbf{b}_{jkl})] \quad (3)$$

and

$$\begin{aligned} V(\theta_{ij}|\mathbf{y}) &= (nd)^{-1} \sum_{k=1}^{n} \sum_{l=d+1}^{2d} (\mathbf{V}_{ij}^{-1} + \psi_{ikl}^{-1})^{-1} \\ &\quad + (nd)^{-1} \sum_{k=1}^{n} \sum_{l=d+1}^{2d} (\mathbf{V}_{ij}^{-1} + \psi_{ikl}^{-1})^{-1} \\ &\quad \times (\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_{ikl}^{-1}(\mathbf{X}_{ij}\alpha_{kl} + \mathbf{Z}_{ij}\mathbf{b}_{jkl})) \\ &\quad \times (\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_{jkl}^{-1}(\mathbf{X}_{ij}\alpha_{kl} + \mathbf{Z}_{ij}\mathbf{b}_{jkl}))^T \\ &\quad \times (\mathbf{V}_{ij}^{-1} + \psi_{jkl}^{-1})^{-1} \\ &\quad - (nd)^{-2} \left\{ \sum_{k=1}^{n} \sum_{l=d+1}^{2d} (\mathbf{V}_{ij}^{-1} + \psi_{jkl}^{-1})^{-1} \right. \\ &\quad \left. \times (\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_{jkl}^{-1}(\mathbf{X}_{ij}\alpha_{kl} + \mathbf{Z}_{ij}\mathbf{b}_{jkl})) \right\} \\ &\quad \times \left\{ \sum_{k=1}^{n} \sum_{l=d+1}^{2d} (\mathbf{V}_{ij}^{-1} + \psi_{jkl}^{-1})^{-1} \right. \\ &\quad \left. \times (\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_{jkl}^{-1}(\mathbf{X}_{ij}\alpha_{kl} + \mathbf{Z}_{ij}\mathbf{b}_{jkl})) \right\}^T. \end{aligned} \quad (4)$$

We use these results in the next section for finding the posterior means and variances of the θ_{ij}'s ($i = 1, \ldots, m; j = 1, \ldots, t; q = 1, \ldots, s$) for special choices of $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$.

3. ESTIMATION OF MEDIAN INCOME OF FOUR-PERSON FAMILIES

The basic data consist of three component vectors $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})^T$ and the associated variance–covariance matrices $\mathbf{V}_{ij}$ ($i = 1, \ldots, 51$; $j = 1, \ldots, 10$). Here Y_{ij1}, Y_{ij2}, and Y_{ij3} are the sample median incomes of four-, three-, and five-person families in state i for year j. The corresponding adjusted census median incomes are denoted by x_{ij1}, x_{ij2}, and x_{ij3}. The true median corresponding to Y_{iju} is denoted by θ_{iju} ($u = 1, 2, 3$). The years $1, \ldots, 10$ correspond to $1980, \ldots, 1989$.

First, consider the trivariate case. Let $\theta_{ij} = (\theta_{ij1}, \theta_{ij2}, \theta_{ij3})^T$. We also write $\mathbf{Y} = (\mathbf{Y}_{11}^T, \ldots, \mathbf{Y}_{51,10}^T)^T$ and $\theta = (\theta_{11}^T, \ldots, \theta_{51,10}^T)^T$. The known design matrices $\mathbf{X}_{ij}$ are given by

$$\mathbf{X}_{ij} = \begin{bmatrix} 1 & x_{ij1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & x_{ij2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{ij3} \end{bmatrix}; \quad i = 1, \ldots, 51, \quad j = 1, \ldots, 10. \quad (5)$$

Also, the vector of regression coefficients is denoted by $\alpha = (\alpha_1, \ldots, \alpha_6)^T$, and the vector of random components for year j is denoted by $\mathbf{b}_j = (b_{j1}, b_{j2}, b_{j3})^T$ ($j = 1, \ldots, 10$). We justify this covariate selection at the end of this section.

We use the HB model of the previous section with $\mathbf{Z}_{ij} = \mathbf{I}_3, \mathbf{S}_j = .00005\mathbf{I}_3$ ($j = 0, 1, \ldots, 10$) and $k_j = 7$ ($j = 0, 1, \ldots, 10$). The choice of the $\mathbf{S}_j$'s and k_j's is to ensure that the posterior distributions are proper, although other choices are clearly possible. Moreover, neither the $\mathbf{S}_j$'s nor the k_j's need to be the same for all j. The main idea behind our choice is to keep the distributions of the hyperparameters nearly diffuse without violating the propriety of the posteriors. Our limited amount of sensitivity analysis of the HB procedure suggests that the choice of hyperpriors does not matter too much when the final goal is to produce posterior means, variances, and covariances of the θ_{ij}'s, the median incomes of four-person families.

A comment about the assumption of known $\mathbf{V}_{ij}$'s also seems in order. Clearly, these are estimates, and ideally, it seems desirable to assign some distributions to the $\mathbf{V}_{ij}$'s of the assumed HB model. However, given the current state of the art, such a task seems near impossible. For instance, if one wanted to assign inverse Wishart priors to the $\mathbf{V}_{ij}$'s,

Table 1. A Comparison of Estimates Under Four Different Criteria

Estimate	Average relative bias	Average squared relative bias	Average absolute bias	Average squared deviation
CPS	.0735	.0084	2,928.82	13,811,122.39
Bureau	.0296	.0013	1,183.90	2,151,350.18
HB^1	.0338	.0018	1,351.67	3,095,736.14
HB^2	.0363	.0021	1,457.47	3,468,496.61
HB^3	.0295	.0013	1,171.71	2,194,553.67
HB^4	.0323	.0016	1,287.78	2,610,249.94
HB^5	.0230	.0009	932.51	1,618,025.33
HB^6	.0295	.0013	1,179.94	2,216,738.06
HB^7	.0287	.0013	1,150.24	2,116,692.71
HB^8	.0324	.0015	1,297.12	2,530,938.06
HB^9	.0271	.0011	1,089.24	1,927,153.24
HB^{10}	.0308	.0014	1,233.59	2,315,875.39

then a formidable task would seem to be the choice of meaningful degrees of freedom. One may add, however, that the V_{ij}'s used are not the raw estimates associated with the Y_{ij}'s, but rather are smoothed versions of the direct CPS estimates and as such are more stable. This estimation procedure, described in detail by Fay et al. (1993, sec. 9.3.2), uses Woodruff's (1952) general approach with appropriate modifications. Even if one does not accept the V_{ij}'s as known, our estimation procedure can genuinely be described as a hierarchical–empirical Bayes approach. The current EB approach of the Bureau of the Census also assumes the V_{ij} as known and also ignores the time series nature of the data.

To implement and monitor the convergence of the Gibbs sampler, we follow the basic approach of Gelman and Rubin (1992). We consider 10 independent sequences each with a sample of size 5,000, and with a burn-in sample of another 5,000.

The implementation requires generation of samples from the full conditionals as given in the Appendix, with one exception. We sample the θ_{ij}'s initially from multivariate t distributions with 2 df having the same location vectors and scale matrices as the corresponding multivariate normal conditionals given in (A.1)–(A.5) of the Appendix. This is based on the Gelman–Rubin idea of initializing certain samples from overdispersed distributions. However, once initialized, the subsequent θ_{ij}'s are sampled from regular multivariate normal conditionals.

To monitor the convergence of the Gibbs sampler, for each $\theta_{i1\delta 1}$ $(i = 1,\ldots,51)$, the ultimate parameters of interest, we follow Gelman and Rubin (1992). Compute $B_{i1\delta 1}/5{,}000$ = the variance between the 10 sequence means $\bar{\theta}_{g1\delta 1}$ each based on 5,000 values; that is, $B_{i1\delta 1}/5{,}000 = \sum_{g=1}^{10}(\bar{\theta}_{gi1\delta 1} - \bar{\theta}_{i1\delta 1})^2/(10-1)$, where $\bar{\theta}_{i1\delta 1} = \sum_{g=1}^{10}\bar{\theta}_{gi1\delta 1}/10$. Also, let $W_{i1\delta 1}$ denote the average of the 10 within-sequence variance, $s^2_{gi1\delta 1}$ each based on $(5{,}000-1)$ df; that is $W_{i1\delta 1} = \sum_{g=1}^{10} S^2_{gi1\delta 1}/10$. Then find

$$\hat{\sigma}^2_{i1\delta 1} = \frac{5{,}000-1}{5{,}000}\,W_{i1\delta 1} + \frac{1}{5{,}000}\,B_{i1\delta 1}$$

Table 2. Percentage Improvements of HB Estimates Over the CPS Estimates Under Four Different Criteria

Estimate	Average relative bias	Average squared relative bias	Average absolute bias	Average squared deviation
HB^1	54.00%	78.18%	53.85%	77.59%
HB^2	50.61%	75.32%	50.24%	74.89%
HB^3	59.91%	84.08%	59.99%	84.11%
HB^4	56.09%	80.98%	56.03%	81.10%
HB^5	68.66%	89.21%	68.16%	88.28%
HB^6	59.84%	84.44%	59.71%	83.95%
HB^7	60.90%	85.02%	60.73%	84.67%
HB^8	55.95%	81.80%	55.71%	81.67%
HB^9	63.11%	86.42%	62.81%	86.05%
HB^{10}	58.13%	83.33%	57.88%	83.23%

and

$$\hat{V}_{i1\delta 1} = \hat{\sigma}^2_{i1\delta 1} + ((10)(5{,}000))^{-1}B_{i1\delta 1}.$$

Finally, find $\hat{R}_{i1\delta 1} = \hat{V}_{i1\delta 1}/\hat{W}_{i1\delta 1}$ $(i = 1,\ldots,51)$. If $\hat{R}_{i1\delta 1}$ $(i = 1,\ldots,51)$ are near 1 for all of the scalar estimands $\hat{\theta}_{i1\delta 1}$ $(i = 1,\ldots,51)$ of interest, then this suggests that the desired convergence is achieved in the Gibbs sampler.

We denote these HB estimates in the trivariate case by HB^9. The corresponding estimates based on a non–time series model utilizes only the census median income figures for 1979, the CPS median income estimates for 1989, and the PCIs for the years 1979 and 1989. In this case, also, we utilize the data available for three-, four-, and five-person families. We denote these estimates by HB^{10}.

Next, we consider several bivariate models, where the basic data for the ith local area is a two-component vector in which the first component is equal to Y_{ij1} and the second component is equal to either Y_{ij2} or Y_{ij3} or $.75Y_{ij2} + .25Y_{ij3}$. Corresponding changes are made in the θ_{ij} vectors, which are now two-component vectors, and the X_{ij} matrices, which are now 2×4 matrices. The resulting HB estimators of the median incomes of four-person families for 1989 are now given by HB^3, HB^5, and HB^7. The corresponding estimates without using a time series model are denoted by HB^4, HB^6, and HB^8.

Finally, in the univariate case only Y_{ij1} $(i = 1,\ldots,51; j = 1,\ldots,10)$ are considered as basic data for estimating

Table 3. Percentage Improvements of HB Estimates Over the Bureau o the Census Estimates Under Four Different Criteria

Estimate	Average relative bias	Average squared relative bias	Average absolute bias	Average squared deviation
HB^1	−14.19%	−40.48%	−14.17%	−43.90%
HB^2	−22.60%	−58.94%	−23.11%	−61.22%
HB^3	−.48%	−2.52%	1.03%	−2.01%
HB^4	− 8.99%	−22.45%	−8.77%	−21.33%
HB^5	22.19%	30.52%	21.23%	24.79%
HB^6	.31%	−.18%	.33%	−3.04%
HB^7	2.94%	3.56%	2.84%	1.61%
HB^8	−9.36%	−17.18%	−9.56%	−17.64%
HB^9	8.42%	12.59%	8.00%	10.42%
HB^{10}	−3.93%	−7.31%	−4.20%	−7.65%

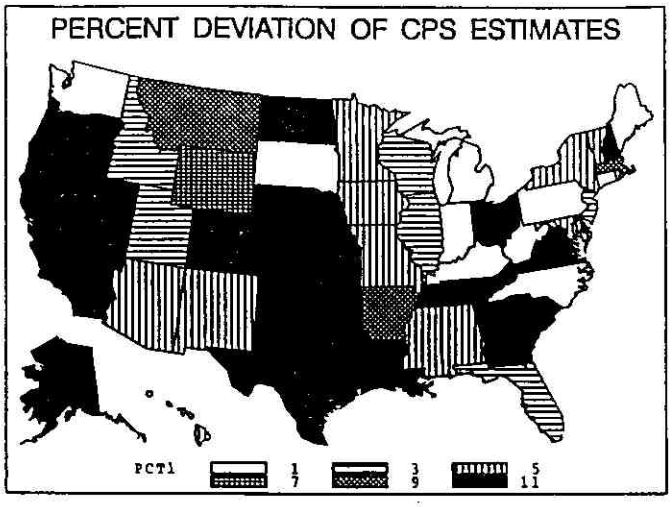

Figure 1. Statewise Deviations (in Percentage) of CPS Median Income Estimates for Four-Person Families From the Corresponding Census Estimates for 1989. The numbers on the right denote midpoints of intervals of length 2; for example, 1 means 0–2%, and so on.

$(\theta_{1,10,1}, \ldots, \theta_{51,10,1})^T$. The $\mathbf{X}_{ij}$ matrices now become two-component row vectors. The resulting HB estimates are denoted by HB^1. For the corresponding model not involving a time series, the HB estimate is denoted by HB^2.

Because all of these estimates are compared to the corresponding census figures, we use the following four criteria to compare the different estimates. Let c_i denote the census estimate for the ith local area ($i = 1, \ldots, 51$). For any estimate $\mathbf{e} = (e_1, \ldots, e_{51})^T$, we compute the following:

- average relative bias $= (51)^{-1} \sum_{i=1}^{51} |c_i - e_i|/c_i$
- average squared relative bias $= (51)^{-1} \sum_{i=1}^{51} |c_i - e_i|^2/c_i^2$
- average absolute bias $= (51)^{-1} \sum_{i=1}^{51} |c_i - e_i|$
- average squared deviation $= (51)^{-1} \sum_{i=1}^{51} (c_i - e_i)^2$.

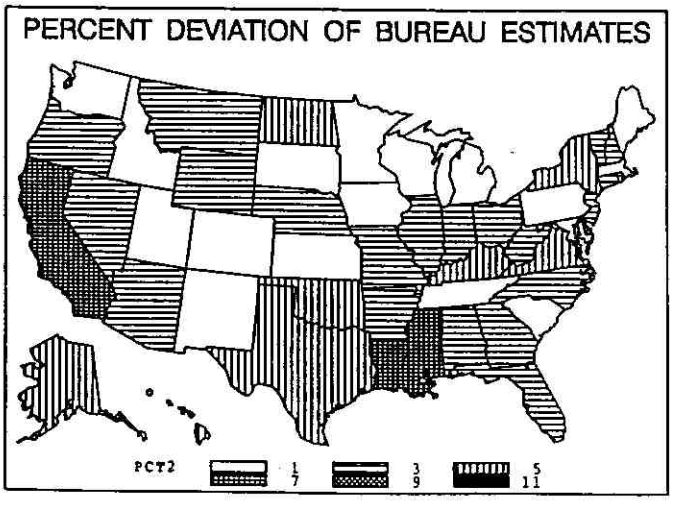

Figure 2. Statewise Deviations (in Percentage) of Census Bureau (EB) Median Income Estimates for Four-Person Families From the Corresponding Census Estimates for 1989. The numbers on the right denote midpoints of intervals of length 2; for example, 1 means 0–2%, and so on.

These four comparison criteria were recommended by the panel on small area estimates of population and income set up by the committee on National Statistics in July 1978, and is available in their July 1980 report (see p. 75).

Table 1 reports these figures for the different estimates. Table 2 gives the percentage improvements over the CPS estimates for 1989. Table 3 presents the corresponding percentage improvements over the estimates prepared by the Bureau of the Census. It is clear from Table 2 that all the HB estimates improve substantially over the CPS estimates according to each one of the four criteria. Moreover HB^5 (i.e., the estimates under the bivariate time series model including the median incomes of four- and five-person families only) seems to work better than the remaining HB estimates. This is much more pronounced in Table 3, where HB^5 improves substantially over the estimates produced by the Bureau of the Census, whereas the remaining HB estimates are either dominated by the Bureau of the Census estimates or improve only moderately over those estimates. To be specific, with the exception of HB^6 (the bivariate non–time series estimates using median incomes of four- and five-person families only), all non–time series estimates perform much worse than the Bureau of the Census estimates, whereas HB^6 is essentially on par with them. With the exceptions of HB^1 and HB^3 (to a certain extent), other time series estimates all improve on the Bureau of the Census estimates, but the best performance comes from HB^5. Indeed, including median income of three-person families seems only to worsen the situation. Thus the trivariate time series estimate HB^9 performs worse than HB^5. The performance declines with HB^7, which includes the weighted average of median incomes of three- and five-person families as the second component variable, with 75% weight attached to the median income of three-person families, and further worsens with HB^3, which includes only the median income of three-person families as the second component variable.

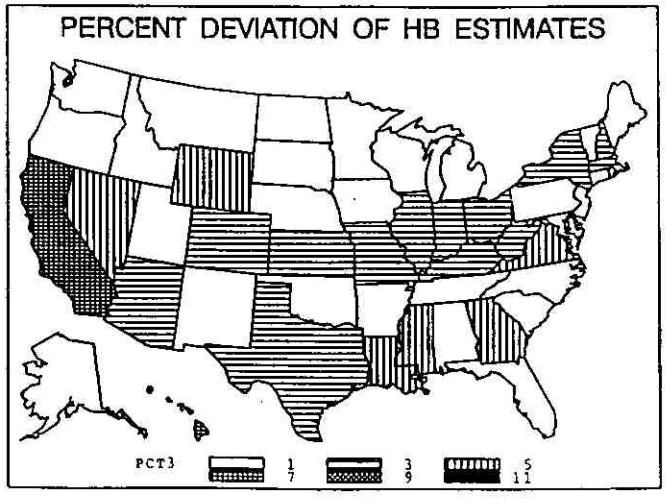

Figure 3. Statewise Deviations (in Percentage) of Optimal BH Median Income Estimates for Four-Person Families From the Corresponding Census Estimates for 1989. The numbers on the right denote midpoints of intervals of length 2; for example, 1 means 0–2%, and so on.

Table 4. A Breakup of the Posterior Variances ($V = V_1 + V_2$) of HB^5 for Selected States

State	V	V_1	V_2
MA	2,205,225	554,000	1,651,225
RI	3,013,696	583,215	2,430,481
OH	1,909,924	680,043	1,229,881
ND	2,295,225	757,625	1,537,600
NE	2,427,364	661,123	1,766,241
TN	2,442,969	606,944	1,836,025
TX	1,633,284	543,348	1,089,936
WY	3,143,529	1,069,929	2,073,600
NV	2,893,401	1,095,120	1,798,281
AK	2,762,244	524,228	2,238,016

Figures 1–3 show statewide percentage deviations of the CPS estimates, Bureau of the Census estimates, and HB^5 estimates from the corresponding estimates as obtained from the 1990 census. Clearly, the CPS estimates have the worst performance. The HB^5 estimates usually perform as well or better than the Bureau of the Census estimates. The performance is particularly better in southern states, including Texas and Florida.

Table 4 provides the posterior variances (V_i) associated with HB^5 for 10 states and gives a breakup of V_i as $V = V_{1i} + V_{2i}$, where

$$V_{1i} = V[E(\theta_{i1\bar{0}1}|\alpha, \mathbf{W}, \mathbf{b}_1, \ldots, \mathbf{b}_{10}, \psi_1, \ldots, \psi_{10}, \mathbf{Y})|\mathbf{Y}]$$

and

$$V_{2i} = E[V(\theta_{i1\bar{0}1}|\alpha, \mathbf{W}, \mathbf{b}_1, \ldots, \mathbf{b}_{10}, \psi_1, \ldots, \psi_{10}, \mathbf{Y})|\mathbf{Y}].$$

This table illustrates that although V_{1i}'s are smaller than V_{2i}'s, ignoring V_{1i}'s in computing the V_i's can lead to underestimation of posterior standard deviations. Thus applying a naive EB method for this problem can often lead inadequate approximations for posterior standard deviations.

Table 5 reports the standard errors SE's associated with the CPS estimates, as well as the different HB estimates once again for 10 states (not the same as those in Table 4). The selection of the states is partially motivated to indicate our findings that although HB^5 and HB^6 have typically the smallest SE's, occasionally other estimates can have similar features. This is evidenced, for example, in the states of Arizona and New York. The Bureau of the Census does not report any SE's associated with their EB estimates. The SE's associated with the HB estimates are always much smaller than the corresponding CPS SE's. Also, the time series estimates usually outperform their non–time series counterparts, the only exception being HB^5 against HB^6, because there does not seem to be a clear-cut winner. However, it is strongly suggested that inclusion of five-person families only along with four-person families leads to a better performance than others.

Finally, Table 6 reports the coefficients of variation associated with the CPS and the different HB estimates. We do not have SE's associated with the estimates produced by the Census Bureau, so coefficient of variation calculation is impossible there.

All the HB methods are far superior to the CPS under this criterion. Once again HB^5 emerges very strong, having 34 coefficients of variation in the 2%–4% range and 17 in the 4%–6% range, but surprisingly, HB^6, the corresponding non–time series estimate, seems to perform even slightly better. At this point, we do not seem to have a very clear explanation of this phenomenon.

Samples from multivariate normal distributions were generated using GASDEV, and the Wishart variables were generated using Bartlett's decomposition. The computations were carried out on a Sun Sparc 10 workstation using Fortran software. The computing time needed to produce all of the tables and graphs was about 8 hours.

The inclusion of only the adjusted census median incomes as covariates in addition to the intercept terms in stage II of the hierarchical model can be justified on two grounds. First, the median income estimates for the 50 states and Washington, D.C., under these models came closest on average to the corresponding census estimates, as compared to models that included only the unadjusted census medians or the saturated models that included both the unadjusted and adjusted census medians. The closeness was decided on the basis of each one of four criteria described earlier. Second, and possibly more important, is a simple model selection device along the lines of Malec and Sedransk (1994). Consider a simple fixed-effects model $\mathbf{Y}_{ij} = \mathbf{X}^*_{ij}\alpha + \mathbf{e}^*_{ij}$, where $\mathbf{e}^*_{ij}$ are independent $N(0, \mathbf{V}_{ij})$. We have different possible choices of $\mathbf{X}^*_{ij}$. Under a specific choice, let $\hat{\mathbf{Y}}_{ij}$ denote the fitted value of $\mathbf{Y}_{ij}$. Then, one computes the deviance $D^2 = \sum_{i=1}^{m}\sum_{j=1}^{t}(\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij})^T\mathbf{V}_{ij}^{-1}(\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij})$, which in the present case turns out to be the weighted sum of squared residuals. Suppose now that M_0, M_1, M_2, and M_3 stand as generic symbols for a saturated model, a model with only the intercept terms, a model with the intercepts and the ad-

Table 5. Estimated Standard Errors for Some Selected States of the Different HB Estimates

States	HB^1	HB^2	HB^3	HB^4	HB^5	HB^6	HB^7	HB^8	HB^9	HB^{10}
ME	2,073	2,016	1,572	1,555	1,525	1,465	1,632	1,602	1,590	1,579
NH	2,328	2,336	1,772	1,881	1,599	1,643	1,779	1,850	1,721	1,790
VT	1,972	1,924	1,565	1,563	1,515	1,473	1,614	1,593	1,580	1,562
NY	1,327	1,310	1,124	1,134	1,060	1,052	1,077	1,098	1,051	1,066
NJ	1,591	1,624	1,293	1,444	1,295	1,409	1,288	1,444	1,256	1,393
IL	1,469	1,448	1,174	1,176	1,136	1,115	1,161	1,164	1,127	1,127
MI	1,369	1,356	1,158	1,172	1,102	1,136	1,150	1,157	1,106	1,115
NM	1,955	1,952	1,452	1,585	1,417	1,459	1,443	1,548	1,417	1,543
AZ	2,405	2,322	1,470	1,714	1,556	1,510	1,752	1,721	1,663	1,652
NV	2,096	2,058	1,659	1,652	1,701	1,702	1,654	1,647	1,611	1,604

Table 6. Coefficient of Variations of Different Estimates

Estimate	Coefficient of variation 2–4%	4–6%	≥6%
CPS	6	7	38*
HB^1	10	37	4
HB^2	10	38	3
HB^3	24	27	0
HB^4	23	28	0
HB^5	34	17	0
HB^6	35	16	0
HB^7	24	27	0
HB^8	22	29	0
HB^9	27	24	0
HB^{10}	26	25	0

* 6–8% 16; ≥8% 22.

justed census medians, and a model with intercepts and unadjusted census medians. The corresponding deviances are denoted by D_0^2, D_1^2, D_2^2, and D_3^2. Clearly, $D_1^2 > D_2^2 > D_0^2$ and $D_1^2 > D_3^2 > D_0^2$. One now computes the ratios (as in Malec and Sedransk 1994)

$$R_2^2 = (D_1^2 - D_2^2)/(D_1^2 - D_0^2)$$

and

$$R_3^2 = (D_1^2 - D_3^2)/(D_1^2 - D_0^2). \tag{6}$$

The ratios R_2^2 and R_3^2 indicate the proportion of the deviation differences between the two extreme models: the saturated model and the intercept model that is captured by the intermediate models.

For the 10 models (identified in the order in which the HB estimates are labeled), we denote the ratios by R_{2i}^2 and R_{3i}^2 $(i = 1, \ldots, 10)$. Clearly, $R_{j(2i-1)}^2 = R_{j(2i)}^2, j = 2, 3; i = 1, \ldots, 5$, because these computations are based only on stages I and II of the hierarchical model of Section 2. Table 7 gives the values of these ratios.

It is clear from Table 7 that the model M_2 which incorporates only the adjusted census medians is the most appropriate model in all the ten cases.

4. SENSITIVITY ANALYSIS

For the particular problem at hand, the assumption of independence of the $\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{it}$ given $\theta_{i1}, \ldots, \theta_{it}$ as made in stage I of the hierarchical model can be legitimately questioned. In the CPS literature, there are strong indications of nontrivial correlation of sampling errors across time. A referee has pointed out that this may be due to primary sampling units common to several surveys. The associate editor suggests that comments of Tiller (1992, p. 152) provide an alternate explanation that the household replacement policy may also be very important and may induce nontrivial correlation patterns that cover several years.

In view of these comments, it seems ideal to have direct empirical assessment of sampling error autocovariances and to include these in stage I of the hierarchical model. Unfortunately, this seems impossible at the moment in the absence of relevent fine-level data, due to confidentiality restrictions. Hence, following the suggestion of the associate

Table 7. Values of the Ratios R_2^2 and R_3^2

i	1, 2	3, 4	5, 6	7, 8	9, 10
R_{2i}^2	.994	.993	.990	.992	.991
R_{3i}^2	.332	.332	.370	.348	.356

editor, we analyze in this section the same data set under alternate models that induce autocorrelations among the $\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{it}$. Specifically, we consider several first-order autoregressive models.

Under a first-order autoregressive model, in stage I of the hierarchical model,

$$\mathbf{Y}_{ij} - \theta_{ij} = \rho(\mathbf{Y}_{i(j-1)} - \theta_{i(j-1)}) + \mathbf{e}_{ij} \qquad (j = 2, \ldots, t), \tag{7}$$

where $\mathbf{e}_{ij}$ are independent $N(0, (1-\rho^2)\mathbf{V}_{ij})$. The factor $1-\rho^2$ is needed to keep $\mathbf{V}(\mathbf{Y}_{ij})$ equal to $\mathbf{V}_{ij}$. Subsequent stages of the HB model are left as before. Then the formulas for the full conditionals given in (A.1)–(A.4) of the Appendix remain unaltered, but (A.5) changes as follows:

- $\theta_{i1}|\mathbf{y}, \alpha, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W}, \theta_{i2}, \ldots, \theta_{it} \overset{\text{ind}}{\sim} N[(\mathbf{V}_{i1}^{-1} + \rho^2\mathbf{V}_{i2}^{-1} + \psi_1^{-1})^{-1}\{\mathbf{V}_{i1}^{-1}\mathbf{y}_{i1} + \rho^2\mathbf{V}_{i2}^{-1}(\mathbf{y}_{i1} + \rho^{-1}(\theta_{i2} - \mathbf{y}_{i2})) + \psi_1^{-1}(\mathbf{X}_{i1}\alpha + \mathbf{Z}_{i1}\mathbf{b}_1)\}, (\mathbf{V}_{i1}^{-1} + \rho^2\mathbf{V}_{i2}^{-1} + \psi_1^{-1})^{-1}]$.
- $\theta_{it}|\mathbf{y}, \alpha, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W}, \theta_{i1}, \ldots, \theta_{i(t-1)} \overset{\text{ind}}{\sim} N[(\mathbf{V}_{it}^{-1} + \psi_t^{-1})^{-1}\{\mathbf{V}_{it}^{-1}(\mathbf{y}_{it} + \rho(\theta_{i(t-1)} - \mathbf{y}_{i(t-1)}) + \psi_t^{-1}(\mathbf{X}_{it}\alpha + \mathbf{Z}_{it}\mathbf{b}_t)\}, (\mathbf{V}_{it}^{-1} + \psi_t^{-1})^{-1}]$.
- For $2 \le j \le t-1$, $\theta_{ij}|\mathbf{y}, \alpha, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W}, \theta_{i1}, \ldots, \theta_{i(j-1)}, \theta_{i(j+1)}, \ldots, \theta_{it} \overset{\text{ind}}{\sim} N[(\mathbf{V}_{ij}^{-1} + \rho^2\mathbf{V}_{i(j+1)}^{-1} + \psi_j^{-1})^{-1}\{\mathbf{V}_{ij}^{-1}(\mathbf{y}_{ij} + \rho(\theta_{i(j-1)} - \mathbf{y}_{i(j-1)}) + \rho^2\mathbf{V}_{i(j+1)}^{-1}(\mathbf{y}_{ij} + \rho^{-1}(\theta_{i(j+1)} - \mathbf{y}_{i(j+1)})) + \psi_j^{-1}(\mathbf{X}_{ij}\alpha + \mathbf{Z}_{ij}\mathbf{b}_j)\}, (\mathbf{V}_{ij}^{-1} + \rho^2\mathbf{V}_{i(j+1)}^{-1} + \psi_j^{-1})^{-1}]$.

We tried the autoregressive models only in the bivariate case that took into account the median income of four- and five-person families only. We considered three different choices of ρ: $\rho = .20, .35$, and $.50$. The resulting HB estimates are denoted by $HB_{.20}^{AR(1)}$, $HB_{.35}^{AR(1)}$, and $HB_{.50}^{AR(1)}$. We compared these estimates to the HB^5 estimates. The results under the alternate models turn out to be quite different from what we obtained under the independence model. Also, a comparison with the census estimates reveals that in the present case, the independence assumption provides better estimates on an average than the corresponding

Table 8. Comparison of HB^5 and $HB^{AR(1)}$

Estimate	Average relative bias	Average squared relative bias	Average absolute bias	Average squared deviation
HB^5	.0230	.0009	932.51	1,618,025.33
$HB_{.20}^{AR(1)}$	.0291	.0013	1,133.45	2,051,716.67
$HB_{.35}^{AR(1)}$	.0279	.0013	1,088.63	1,956,516.51
$HB_{.50}^{AR(1)}$	.0274	.0012	1,079.73	1,905,051.53

first-order autoregressive assumption. This is documented in Table 8.

Needless to say, we could have considered other autoregressive models in which the HB estimates could have performed better. The objective of this section is only to demonstrate that estimates can be different under different first-stage assumptions.

5. CONCLUDING REMARKS

This article has presented hierarchical Bayes times series modeling for estimating the median income of four-person families in the 50 states and the District of Columbia. A comparison of these estimates with those obtained from the 1990 decennial census reveals that a bivariate time series model utilizing the median incomes of four- and five-person families performs the best and is clearly an attractive alternative to the existing methodology of the Bureau of the Census. Second, the HB approach has the additional advantage of providing standard errors along with the point estimates. The authors have software available for implementing the proposed methodology. The data and the Fortran codes are given in STATLIB.

APPENDIX: FORMULAS FOR THE FULL CONDITIONALS

- $\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W} \sim N_p[(\sum_{i=1}^{m}\sum_{j=1}^{t}\mathbf{X}_{ij}^T \psi_j^{-1}\mathbf{X}_{ij})^{-1}\sum_{i=1}^{m}\sum_{j=1}^{t}\mathbf{X}_{ij}^T\psi_j^{-1}(\theta_{ij} - \mathbf{Z}_{ij}^T\mathbf{b}_j), (\sum_{i=1}^{m}\sum_{j=1}^{t}\mathbf{X}_{ij}^T\psi_j^{-1}\mathbf{X}_{ij})^{-1}]$. (A.1)
- For $j = 2, \ldots, t-1$, $\mathbf{b}_j|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_{j-1}, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W} \sim N_q[(\sum_{i=1}^{m}\mathbf{Z}_{ij}^T\psi_j^{-1}\mathbf{Z}_{ij} + 2\mathbf{W}^{-1})^{-1}(\sum_{i=1}^{m}\mathbf{Z}_{ij}^T\psi_j^{-1}(\theta_{ij} - \mathbf{X}_{ij}\alpha) + \mathbf{W}^{-1}(\mathbf{b}_{j-1} + \mathbf{b}_{j+1})), (\sum_{i=1}^{m}\mathbf{Z}_{ij}^T\psi_j^{-1}\mathbf{Z}_{ij} + 2\mathbf{W}^{-1})^{-1}]$. (A.2a)
- For $j = 1$, $\mathbf{b}_1|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{b}_2, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W} \sim N_q[(\sum_{i=1}^{m}\mathbf{Z}_{i1}^T\psi_1^{-1}\mathbf{Z}_{i1} + \mathbf{W}^{-1})^{-1}(\sum_{i=1}^{m}\mathbf{Z}_{i1}^T\psi_1^{-1}(\theta_{i1} - \mathbf{X}_{i1}\alpha) + \mathbf{W}^{-1}\mathbf{b}_2), (\sum_{i=1}^{m}\mathbf{Z}_{i1}^T\psi_1^{-1}\mathbf{Z}_{i1} + \mathbf{W}^{-1})^{-1}]$. (A.2b)
- For $j = t$, $\mathbf{b}_t|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_{t-1}, \psi_1, \ldots, \psi_t, \mathbf{W} \sim N_q[(\sum_{i=1}^{m}\mathbf{Z}_{it}^T\psi_t^{-1}\mathbf{Z}_{it} + \mathbf{W}^{-1})^{-1}(\sum_{i=1}^{m}\mathbf{Z}_{it}^T\psi_t^{-1}(\theta_{it} - \mathbf{X}_{it}\alpha) + \mathbf{W}^{-1}\mathbf{b}_{t-1}), (\sum_{i=1}^{m}\mathbf{Z}_{it}^T\psi_t^{-1}\mathbf{Z}_{it} + \mathbf{W}^{-1})^{-1}]$. (A.2c)
- $\psi_j^{-1}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_t, \mathbf{W} \overset{\text{ind}}{\sim} \text{Wishart}[\mathbf{S}_j + \sum_{i=1}^{m}(\theta_{ij} - \mathbf{X}_{ij}\alpha - \mathbf{Z}_{ij}\mathbf{b}_j)(\theta_{ij} - \mathbf{X}_{ij}\alpha - \mathbf{Z}_{ij}\mathbf{b}_j)^T, k_j + m]$. (A.3)
- $\mathbf{W}^{-1}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t \sim \text{Wishart}[\mathbf{S}_0 + \sum_{j=1}^{t}(\mathbf{b}_j - \mathbf{b}_{j-1})(\mathbf{b}_j - \mathbf{b}_{j-1})^T, k_0 + t]$, where $\mathbf{b}_0 = \mathbf{0}$. (A.4)
- $\theta_{ij}|\mathbf{y}, \boldsymbol{\alpha}, \mathbf{b}_1, \ldots, \mathbf{b}_t, \psi_1, \ldots, \psi_t, \mathbf{W} \overset{\text{ind}}{\sim} N[(\mathbf{V}_{ij}^{-1} + \psi_j^{-1})^{-1}(\mathbf{V}_{ij}^{-1}\mathbf{y}_{ij} + \psi_j^{-1}(\mathbf{X}_{ij}\alpha + \mathbf{Z}_{ij}\mathbf{b}_j)), (\mathbf{V}_{ij}^{-1} + \psi_j^{-1})^{-1}]$. (A.5)

[Received February 1993. Revised May 1996.]

REFERENCES

Bell, W. (1984), "Signal Extraction for Nonstationary Time Series," *The Annals of Statistics*, 12, 646–664.

Broemeling, L. (1985), *Bayesian Analysis of Linear Models*, New York: Marcel Dekker.

Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), "Estimation of Median Income of Four-Person Families: A Bayesian Approach," in *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zeller*, eds. D. A. Berry, K. M. Chaloner, and J. K. Geweke, New York: Wiley, pp. 129–140.

Fay, R. E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in *Small Area Statistics*, eds. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, New York: Wiley, pp. 91–102.

Fay, R. E., Nelson, C. T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21, Washington, DC: Statistical Policy Office, Office of Management and Budget, pp. 901–917.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

——— (1991), "Gibbs Sampling for Marginal Posterior Expectations," *Communications in Statistics, Part A—Theory and Methods*, 20, 1747–1766.

Gelman, A. E., and Rubin, D. (1992), "Inference From Iterative Simulation" (with discussion), *Statistical Science*, 7, 457–511.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Malec, D., and Sedransk, J. (1994), "Small Area Inference for Binary Variables in the National Health Interview Survey," preprint.

Meinhold, R., and Singpurwalla, N. D. (1983), "Understanding the Kalman Filter," *The American Statistician*, 37, 123–137.

National Research Council (1980), *Panel on Small Area Estimates of Population and Income. Estimating Population and Income of Small Areas*, Washington, DC: National Academic Press.

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

West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting, *Journal of the American Statistical Association*, 80, 73–97.

Woodruff, R. (1952), "Confidence Intervals for Medians and Other Position Measures," *Journal of the American Statistical Association*, 47, 635–646.

Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data

ROBERT E. FAY III and ROGER A. HERRIOT*

An adaptation of the James-Stein estimator is applied to sample estimates of income for small places (i.e., population less than 1,000) from the 1970 Census of Population and Housing. The adaptation incorporates linear regression in the context of unequal variances. Evidence is presented that the resulting estimates have smaller average error than either the sample estimates or an alternate procedure of using county averages. The new estimates for these small places now form the basis for the Census Bureau's updated estimates of per capita income for the General Revenue Sharing Program.

KEY WORDS: Biased estimation; Small-area statistics; James-Stein; Income; Revenue sharing.

1. INTRODUCTION

The State and Local Fiscal Assistance Act of 1972 specifies the distribution of funds to states and units of general-purpose local government for operational or capital expenditures. The resulting General Revenue Sharing Program, administered by the Treasury Department, allocates monies to state and local governments on the basis of interdependent formulas: Funds are distributed to approximately 39,000 units of local government by dividing state allocations. Statistics on population, per capita income (PCI), and adjusted taxes are used to determine the allocations within states.

The Census Bureau provides the Treasury Department with current estimates of these statistics for the states and local jurisdictions receiving funds under the General Revenue Sharing Program. Separate methodologies are used to update the population counts and the income figures from the 1970 Census of Population and Housing. Data from the Internal Revenue Service (IRS) and the Bureau of Economic Analysis form the basis for updating the census estimates of income. In general, the 1970 census values of PCI in 1969 are multiplied by ratios of an administrative estimate of PCI in the current year and a similarly derived estimate for 1969. Herriot (1977) described this methodology in greater detail.

The 1970 census thus constitutes the foundation for the current estimates of PCI, but a significant problem arises in this regard. Of the estimates required, more than one-third, or approximately 15,000, are for places with population of fewer than 500 persons in 1970. Because income was collected on the basis of a 20 percent sample in the 1970 census, the sampling error for the estimates for such small places is an important consideration. For a place of 500 persons, the coefficient of variation (relative standard error) for the 1970 census estimate of PCI is about 13 percent; for a place of 100 persons, 30 percent. The magnitude of these sampling errors initially led the Census Bureau and the Treasury Department to agree to set aside the census figures for these places and to substitute the respective county average figures instead.

This substitution of the county figures for the census estimates for places with fewer than 500 persons would seem to be based on the following statistical reasoning: For larger places the sampling errors of the census sample estimates are sufficiently small so that they might be chosen as the best estimates, but for smaller places substituting biased estimates with negligible sampling error (the county values) for estimates with large sampling error is preferable. This sort of reasoning is of course present, formally or informally, in a great deal of statistical practice. Aspects of this particular problem suggested, however, that this initial solution might be improved considerably: The dividing line of 500 persons was essentially an arbitrary choice; the census estimates for a significant number of small places were many standard errors removed from the county values that had been substituted, thus suggesting a failure of the county values to represent adequately the true values for these places; and auxiliary data related to PCI from the IRS and the 1970 census had not been incorporated in the estimation. In this article we shall describe the application of procedures adapted from the original James-Stein estimator to the problem of estimating 1969 PCI for these small places by addressing each of the deficiencies of the original choice. The revised estimator consisted of the following elements:

1. Fitting a regression equation to the census sample estimates, using as independent variables the

* Robert E. Fay III is a Staff Assistant, Statistical Methods Division, and Roger A. Herriot is Assistant Chief for Socio-economic Programs, Population Division, both at the U.S. Bureau of the Census, Washington, DC 20233. The authors wish to thank staff members, particularly Emmet Spiers, for their help on the project and to acknowledge the support of Daniel B. Levine, Associate Director for Demographic Fields, and of Harold Nisselson, Associate Director for Statistical Standards and Methodology. The authors also wish to thank Carl Morris for helpful comments on the research and to express appreciation to an associate editor and a referee for their suggestions on the exposition of this material.

© Journal of the American Statistical Association
June 1979, Volume 74, Number 366
Applications Section

county values, tax-return data for 1969, and data on housing from the 1970 census;

2. Measuring the goodness of fit between the regression equation and the sample data, taking into consideration the expected contribution of sampling error to the observed differences, and deriving an estimated measure of average lack of fit between the regression estimates and the underlying true values for the places;
3. Forming a weighted average of the sample and the regression estimate for each place, adjusting the weights to reflect the relative magnitudes of the average lack of fit of the regression and the variance of the sample estimate; and
4. Constraining each such weighted average to be within one standard error of the sample estimate, thus preventing severe disagreement between the sample and final estimate.

Because of the mathematical and logical consistency of the revised procedures, and on the basis of independent empirical evidence, the Census Bureau has used this methodology in forming the estimates for 1974 and subsequent years. To our knowledge, the Census Bureau's use is the largest application of James-Stein procedures in a federal statistical program.

2. THE JAMES-STEIN ESTIMATOR AND ITS DESCENDANTS

In order to describe the nature of the estimator that we developed for this problem, we will briefly review some of its predecessors. Other authors, for example, Efron and Morris (1973a, 1975), have given a more comprehensive presentation of much of the material summarized in this section.

Suppose that we have a single observation $\mathbf{Y} = (Y_1, \ldots, Y_k)^T$ from a k-dimensional multivariate normal distribution with mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ and covariance matrix $D\mathbf{I}$, where D is a known scalar constant. Equivalently, the Y_i's are assumed to be independent and identically distributed according to normal distributions with means θ_i and variance D, that is, $Y_i \sim_{\text{ind}} N(\theta_i, D)$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is $\mathbf{Y}$; each Y_i is the obvious estimate of its respective θ_i. Stein (1955) showed that for $k \geq 3$, $\mathbf{Y}$ is not admissible under the usual loss function defined for an estimator $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)^T$ by

$$R(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = EL(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_i E_{\theta_i}(\theta_i - \theta_i^*)^2 . \quad (2.1)$$

We have, of course, $R(\boldsymbol{\theta}, \mathbf{Y}) = kD$. For $k \geq 3$, James and Stein (1961) exhibited the estimator $\boldsymbol{\delta}' = (\delta_1', \ldots, \delta_k')^T$ defined by

$$\delta_i' = (1 - ((k-2)D/S))Y_i , \quad (2.2)$$

where

$$S = \sum_i Y_i^2 \quad (2.3)$$

with risk $R(\boldsymbol{\theta}, \boldsymbol{\delta}') < kD$ for all $\boldsymbol{\theta}$. Consequently, $\boldsymbol{\delta}'$ dominates the maximum likelihood estimator $\mathbf{Y}$ with respect to the loss function (2.1).

The result is far from obvious: The Y_i's estimate the respective θ_i's, which in turn need to have no specific relationship to each other; yet by combining information from apparently unrelated estimation problems, the expected total loss (2.1) may be reduced. To do this, $\boldsymbol{\delta}'$ in effect shrinks $\mathbf{Y}$ towards $\mathbf{0}$; that is, each component of $\mathbf{Y}$ is proportionally reduced by the same factor. The amount of shrinkage depends on the relative closeness of $\mathbf{Y}$ to $\mathbf{0}$; for $\mathbf{Y}$ near $\mathbf{0}$, the shrinkage is substantial, while for $\mathbf{Y}$ far from $\mathbf{0}$, $\boldsymbol{\delta}'$ becomes essentially $\mathbf{Y}$. Roughly speaking, to the extent that $\boldsymbol{\theta}$ lies close to $\mathbf{0}$, $\mathbf{Y}$ is also in a sense an estimate of $\mathbf{0}$, and $\boldsymbol{\delta}'$ incorporates this information in estimating $\boldsymbol{\theta}$.

James and Stein noted that (2.2) could be uniformly improved for all $\boldsymbol{\theta}$ by restricting $(k-2)D/S$ to $[0, 1]$, replacing this term by 1 in cases in which it was greater. This restriction prevents $\mathbf{Y}$ from being partially reflected through the origin and is routinely incorporated in applications of (2.2).

The estimator $\boldsymbol{\delta}'$ has inspired a number of important variations. The link between $\boldsymbol{\delta}'$ and many of the subsequent adaptations can be traced most easily through the correspondence between (2.2) and a classical Bayes estimator. Suppose that we assume that $\boldsymbol{\theta}$ has a prior distribution $\theta_i \sim_{\text{ind}} N(0, A)$, that is, normal with variance A. Then the Bayes estimator $\boldsymbol{\theta}_B^*$ of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}_B^* = (1 - (D/(A + D)))\mathbf{Y} . \quad (2.4)$$

Thus, the Bayes estimator in this situation also shrinks $\mathbf{Y}$ towards $\mathbf{0}$.

The James-Stein estimator (2.2) mimics the Bayes estimator in the following manner: Under the given prior distribution $\theta_i \sim_{\text{ind}} N(0, A)$, the expectation of $(k-2)D/S$, taken over the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{Y}$, is $D/(D + A)$, showing the correspondence between (2.2) and (2.4). In the Bayesian context, regardless of the value of A, (2.2) approximates the Bayes estimator (2.4) by in effect estimating A on the basis of $\mathbf{Y}$. This principle forms the basis from which the other estimators discussed here are derived. In each instance, an estimate A^* of A is obtained, providing both a notion of the average variation of θ_i about some prior estimate and an indication of how much weight should be given to the prior and sample estimates in order to estimate θ_i.

An immediate generalization of (2.2) follows in the case in which a p-dimensional row vector $\mathbf{X}_i$ is available for each θ_i, representing auxiliary information about θ_i. For $Y_i \sim_{\text{ind}} N(\theta_i, D)$ and $\theta_i \sim_{\text{ind}} N(\mathbf{X}_i\boldsymbol{\beta}, A)$, with a uniform (improper) prior distribution on $\boldsymbol{\beta}$, the regression estimate (for $\mathbf{X}^T\mathbf{X}$ of full rank p),

$$Y_i^* = \mathbf{X}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \quad (2.5)$$

may be combined with the sample estimate Y_i to form the Bayes estimator

$$\theta_{Bi}^* = Y_i^* + (1 - (D/(D + A)))(Y_i - Y_i^*) \quad (2.6)$$

in a form similar to (2.4). Equivalently,

$$\theta_{Bi}^* = (D/(D + A))Y_i^* + (A/(D + A))\tilde{Y}_i \quad (2.7)$$

expresses the estimator as a weighted average of Y_i^* and Y_i. The James-Stein analogue of (2.6) and (2.7) for $p < k - 2$ is

$$\delta_i' = Y_i^* + (1 - ((k - p - 2)D/S))(Y_i - Y_i^*) \quad (2.8)$$

$$= ((k - p - 2)D/S)Y_i^* + (1 - ((k - p - 2)D/S))Y_i \quad (2.9)$$

where

$$S = \sum_i (Y_i - Y_i^*)^2 . \quad (2.10)$$

A special case of (2.5) and (2.8) through (2.10) is for $p = 1$, $X_i = 1$: (2.5) makes each Y_i^* the mean of all of the Y_i's, and (2.9) averages each Y_i with the mean. (The estimator (2.5) and (2.8) through (2.10) in the general case in fact follows directly from (2.2) and (2.3) without requiring a Bayesian formulation, but the intent of the estimator is more clearly illustrated in the Bayesian context.)

Efron and Morris (1971, 1972) remarked that both Bayes estimators such as (2.4) and empirical Bayes estimators such as (2.2) may perform well overall but poorly on individual components. In these instances the shrinkage of (2.2) or (2.4), which benefits most components of Y, is singularly inappropriate for the particular θ_i. For the Bayes case (2.4), θ_i may be unusual relative to the prior distribution, while for the empirical Bayes case (2.2), θ_i may lie much further from 0 than the other components of θ. Efron and Morris suggested a straightforward compromise, which consists of restricting the amount by which δ_i' differs from Y_i by some multiple of the standard error of Y_i. With this restriction, (2.2) becomes

$$\delta_i'' = \delta_i' \quad \text{if} \quad Y_i - c \le \delta_i' \le Y_i + c \quad (2.11)$$

$$= Y_i - c \quad \text{if} \quad \delta_i' < Y_i - c \quad (2.12)$$

$$= Y_i + c \quad \text{if} \quad \delta_i' > Y_i + c \quad (2.13)$$

The estimator (2.11) through (2.13) compromises between limiting the maximum possible risk to any component and preserving the average gains of δ'. The choice $c = D^{\frac{1}{2}}$, for example, ensures that $E(\delta_i'' - \theta_i)^2 < 2D$, while retaining more than 80 percent of the average gain of δ' over Y.

For $Y_i \sim_{\text{ind}} N(\theta_i, D_i)$, the possible strategies for extending the James-Stein estimator are numerous but more theoretically difficult if the D_i's are known but not all equal. The most simple extension of (2.2) may be derived by assuming a Bayes prior $\theta_i \sim_{\text{ind}} N(0, AD_i)$. This problem may be solved by transforming Y, applying (2.2) to the vector of elements $Y_i/D_i^{\frac{1}{2}}$, which have the common variance $D = 1$. The resulting δ' from (2.2) may be transformed back to the original scale by computing $\delta_i' D_i^{\frac{1}{2}}$. Even outside the Bayesian formulation, this estimator dominates the maximum likelihood estimator Y with respect to the loss function

$$R(\theta, \hat{\theta}) = \sum_i E_{\theta_i}(\theta_i - \theta_i^*)^2/D_i \quad (2.14)$$

for all θ. (A similar approach may be used to extend (2.8).) This estimator will be most effective against a Bayes prior in which the variance of the prior distribution is proportional to the sampling variance. The resulting estimator applies an equal amount of shrinkage to each component of Y.

In many applications, however, the linkage between the sampling variance of Y_i about θ_i and the Bayes variance of θ_i about 0 is less direct. An alternate approach is to develop an estimator that more closely parallels the Bayes estimator for the prior distribution $\theta_i \sim_{\text{ind}} N(0, A)$, that is, with constant prior variance regardless of D_i. Efron and Morris (1973a) first proposed an extension of (2.2) under this second assumption. The estimator that we used in this application, however, more closely resembled one suggested by Carter and Rolph (1974). In considering the situation $Y_i \sim_{\text{ind}} N(\theta_i, D_i)$ and $\theta_i \sim_{\text{ind}} N(\nu, A)$, with known D_i but unknown ν and θ_i, they observed for the weighted sample mean

$$\nu^* = \sum_i Y_i/(A + D_i)/\sum_i 1/(A + D_i) \quad (2.15)$$

that

$$E\left(\sum_i \frac{(Y_i - \nu^*)^2}{A + D_i}\right) = k - 1 \quad (2.16)$$

for the joint expectation over both Y and θ, when A is a known constant. They suggested estimating A as the unique solution $A^* \ge 0$ such that (2.15) and (2.16) are simultaneously satisfied when the expectation operator is omitted from (2.16),

$$\sum_i \frac{(Y_i - \nu^*)^2}{A^* + D_i} = k - 1 . \quad (2.17)$$

They set $A^* = 0$ if no positive joint solution of (2.15) and (2.17) exists. Each θ_i is estimated by a weighted average of Y_i and ν^*,

$$\delta_i' = (A^*/(A^* + D_i))Y_i + (D_i/(A^* + D_i))\nu^* . \quad (2.18)$$

The estimator that we applied to the 1970 census estimates of PCI is an extension of (2.15), (2.17), and (2.18) to the linear regression case. We considered $Y_i \sim_{\text{ind}} N(\theta_i, D_i)$ and $\theta_i \sim_{\text{ind}} N(\mathbf{X}_i\beta, A)$ for a p-dimensional row vector $\mathbf{X}_i$ and regression coefficients β with an (improper) uniform prior distribution. The row vectors $\mathbf{X}_i$ and sampling variances D_i were known, but β and A were both to be estimated from the data.

To derive the estimator, we first considered relationships when A was known. Over the joint distribution of Y and θ in this case, the weighted regression estimates

$$Y_i^* = \mathbf{X}_i(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \quad (2.19)$$

where $\mathbf{V}$ is a diagonal matrix with $V_{ii} = D_i + A$ give the minimum variance unbiased estimates of $\mathbf{X}_i\beta$, the prior means of θ_i. (These estimates are also the posterior means of $\mathbf{X}_i\beta$.) Over this same joint distribution with

known A,

$$E\left(\sum_i \frac{(Y_i - Y_i^*)^2}{A + D_i}\right) = k - p \ . \qquad (2.20)$$

Equation (2.20) is a standard result in weighted least squares under the preceding assumptions and may be found in texts by Rao (1965, pp. 187–188) and by Draper and Smith (1966, pp. 77–81).

Following the program of Carter and Rolph, we estimated A from the data by removing the expectation operator from (2.20)

$$\sum_i \frac{(Y_i - Y_i^*)^2}{A^* + D_i} = k - p \qquad (2.21)$$

and found the unique $A^* \geq 0$ solving both (2.21) and (2.19), using $A^* = 0$ when no positive solution could be found. The estimator was then

$$\delta_i' = (A^*/(A^* + D_i))Y_i + (D_i/(A^* + D_i))Y_i^* \ . \qquad (2.22)$$

This weighted average of the sample and regression estimate would be the classical Bayes estimator in the case that A were known. The restrictions (2.11) through (2.13) were then imposed on each component δ_i' with $c = D_i^{\frac{1}{2}}$. The actual numerical operations used to solve equations (2.19) and (2.21) simultaneously are described in the Appendix. (We also discuss there an alternate estimator for this problem based on a maximum likelihood approach to fitting the model $\theta_i \sim_{\text{ind}} N(\mathbf{X}_i\beta, AD_i^\alpha)$, where β, A, and α may be jointly estimated from the data.)

We have traced the development of this estimator here through its relation to general results for the James-Stein estimator; yet parallel research in estimation for local areas also precedes these results. Ericksen's work (1973, 1974) explored use of sample data to determine regression estimates for small areas, and Madow (see Madow and Hansen 1975) first remarked on the merit of forming a weighted average of the sample and regression estimates. The estimator presented here represents a further development of these basic ideas.

3. APPLICATION TO ESTIMATION OF INCOME FOR SMALL PLACES

This section will describe the steps used to apply the preceding theory to the estimation of PCI in 1969. The elements of the approach consisted of

1. A division of the total problem into a set of separate estimation problems;
2. Logarithmic transformation of the census values to a scale in which the sampling variances could be considered known;
3. Identification and similar transformation of auxiliary variables available for each place;
4. Derivation of a regression estimate for each place, which was combined with the sample estimate by using (2.19), (2.21) and (2.22), and (2.11) through (2.13);
5. Retransformation of the resulting estimates back to the original scale; and
6. A final proportional readjustment of the resulting estimates to sum to sample estimates of total income at the state and county level.

The following discussion treats each of these points in detail.

Although the initial substitution of the county values of PCI had been carried out only for places of population less than 500 before this investigation, we extended the problem to all places with 20 percent sample estimates of population less than 1,000. (The 20 percent sample count, which is approximately proportional to the number of sample persons in the place, is often in minor disagreement with the complete count for places of this size.) We divided the overall problem into 100 separate estimation problems along two dimensions: a division between places with 20 percent sample estimates of population less than 500 and those between 500 and 999, and an independent consideration of each state. An average of 200 to 300 places with population less than 500 in a given state were thus treated as a joint estimation problem, although there was considerable variation in the size of this group. In some states only 10 or 20 cases were involved. (In addition, some states required estimates for two kinds of geography, places and townships. For simplicity we will discuss the problem for places only, although parallel procedures were applied separately to obtain estimates for the townships.)

For almost all places, a sample estimate Z_i and a weighted 20 percent sample count N_i were available. As part of the processing of the 1970 census, variance computations were performed in eight states and the findings generalized to the rest of the country (U.S. Bureau of the Census 1976, pp. 11-8–11-9). An unpublished finding of this generalization was the approximation of the coefficient of variation of Z_i as $3.0/N_i^{\frac{1}{2}}$. Because the coefficient of variation does not depend on the expected value, the standard deviation increases in direct proportion to the expected value. Hence, the log transformation stabilizes this variance, and the variance of $Y_i = \ln(Z_i)$, the natural logarithm of Z_i, is approximately $9.0/N_i$ and does not depend on the expected value of Z_i. This procedure of stabilizing the variances has appeared in some other applications of the James-Stein estimator (e.g., Carter and Rolph 1974).

Each place, without exception, has an associated county value of PCI from the 1970 census. (With a handful of exceptions, places do not cross county lines.) We computed the natural logarithms of these county figures for use as an independent variable in the regression model. Because of the considerably larger county populations, this variable has typically negligible sampling error.

Two other important sources of data are available for these places: the value of owner-occupied housing from the 1970 census, and the average adjusted gross income

per exemption from the 1969 IRS returns for 1969. Both variables are free from sampling error, but each has other limitations. The value of owner-occupied housing was collected in the 1970 census only for nonfarm dwellings; we consequently chose to omit this variable from the analysis for places with a substantial proportion of farm residences. The IRS data, on the other hand, are affected by errors in coding tax returns to Census Bureau geography on the basis of mailing address. Some places more than others are affected by substantial ambiguity between the mailing addresses and place boundaries. Places thus affected were identified on the basis of unusual ratios between the number of exemptions coded to the place and the 100 percent population count, and in such cases the IRS results were omitted from the analysis. The IRS results were also dropped for places with significant boundary changes since 1970.

After editing the IRS and housing data in the preceding fashion, the natural logarithms of each of these variables were taken, whenever the case met the criteria for inclusion, and matched to logarithms of the respective county values for these variables. Four separate regressions were possible:

1. A constant term and the logarithm of PCI for the county (with $p = 2$ in the notation of the preceding section);
2. A constant term, the logarithm of PCI for the county, and logarithms of the value of housing for both the place and the county ($p = 4$);
3. A constant term, the logarithm of PCI for the county, and logarithms of IRS-adjusted gross income per exemption for both the place and the county ($p = 4$); and
4. A constant term, the logarithm of PCI for the county, the logarithms of the value of housing for both the place and the county, and the logarithms of IRS-adjusted gross income per exemption for both the place and the county ($p = 6$).

Inclusion of both the logarithms of the county and place values for either the housing or tax data is mathematically equivalent to inclusion of both the logarithm of the place value and the logarithm of the ratio of the place to county values. Thus, the regression was able to use the data for the places on an absolute scale, across the entire state, and in relation to the county values.

Our strategy consisted of computing each of the four regressions for those Y_i's with the necessary independent variables for the particular regression by solving (2.19) and (2.21). Using the regression equation corresponding to all the available variables for each place, we computed (2.22) subject to a constraint of the form (2.11) through (2.13). For states with only a few small places, the number of regressions fitted was restricted by insufficient data. Places without any census sample estimate were estimated directly from the regression (2.19).

The preceding estimates developed on the logarithmic scale were transformed back to the original scale. A final two-dimensional iterative proportional adjustment (raking) was applied to all places in each state, including those with population more than 1,000, to force two constraints: the addition of total estimated income (PCI times population) for places belonging to the classes of places with less than 500, 500 to 999, and more than 1,000, to the sample estimates of these totals at the state level; and the addition of the estimates for all places, disregarding size, within a county to the sample estimate of the total for all places in the county. These adjustments, on the order of 1 or 2 percent, were quite small relative to the other aspects of this estimation problem, but they imposed a logical consistency on the outcome and ensured that the analysis of the data on the logarithmic scale did not induce systematic bias across all small places.

The values of A^* provide a measure of the average fit of the regression models to the sample data, after allowance is made for sampling error in Y_i. Table 1 shows the values of A^* obtained for the states with the largest number of places of size less than 500. In a sense, a value for A^* of .045 indicates an average level of accuracy equivalent to the accuracy of a sample estimate for a place of size 200 (9.0/200 = .045, from the formula for the approximate coefficient of variation noted earlier), because (2.22) weights the sample and regression estimates equally in this case. (The value .045 for A^* may be thought to correspond to an average—in the sense of root mean square—error of prediction by the regression of the true value of PCI of approximately 21 percent, because $.045 = 0.21^2$.) In turn, the expected improve-

1. Estimated A^* for Places With 20 Percent Sample Estimates of Population Less Than 500

States	Regression Equation: County	County and Tax	County and Housing	County, Tax, and Housing
States With More Than 500 Places in Class				
Illinois	.036	.032	.019	.017
Iowa	.029	.011	.017	.000
Kansas	.064	.048	.016	.020
Minnesota	.063	.055	.014	.019
Missouri	.061	.033	.034	.017
Nebraska	.065	.041	.019	.000
North Dakota	.072	.081	.020	.004
South Dakota	.138	.138	.014	—
Wisconsin	.042	.025	.025	.004
States With 200 to 500 Places in Class				
Arkansas	.074	.036	.039	.018
Georgia	.056	.081	.067	.114
Indiana	.040	.012	.003	.000
Maine	.052	.015	—	—
Michigan	.040	.032	.028	.023
Ohio	.034	.015	.004	.004
Oklahoma	.063	.027	.049	.036
Pennsylvania	.020	.018	.016	.011
Texas	.092	.048	.056	.040

NOTE: A dash (—) indicates that the regression was not fitted because of too few observations.

2. Estimated A^* for Places With 20 Percent Sample Estimates of Population 500 to 999

States	Regression Equation: County	County and Tax	County and Housing	County, Tax, and Housing
States With More Than 250 Places in Class				
Illinois	.032	.023	.012	.008
Indiana	.017	.014	.007	.009
Michigan	.019	.014	.005	.008
Minnesota	.056	.040	.021	.007
New York	.052	.015	.028	.006
Ohio	.024	.010	.005	.000
Pennsylvania	.035	.025	.015	.026
Wisconsin	.039	.030	.014	—
States With 100 to 250 Places in Class				
Iowa	.017	.005	.016	.004
Kansas	.025	.010	.014	.008
Maine	.022	.021	—	—
Missouri	.042	.019	.011	.013
Nebraska	.027	.007	.008	.008
Texas	.050	.017	.013	.012

NOTE: A dash (—) indicates that the regression was not fitted because of too few observations.

ment of an equal weighting of two estimates with equivalent estimates of error would be to reduce the variance by one-half, or to give, on the average, the combined estimate an accuracy that would be achieved by a sample estimate alone for a place of 400 persons, that is, a relative error of about 15 percent.

In fact, for the regression equation based on county values alone, more than half the states in Table 1 have values for A^* greater than .045, suggesting that the county value is often not so good a prediction of the true value as the sample estimate for places with more than 200. Parenthetically, this finding suggests that the original decision that had preceded this investigation, namely, to replace the sample estimates with the county values for places of size less than 500, actually exaggerated the ability of the county values alone to serve as a good prediction for these places. If no James-Stein estimation was to have been done, it would have been better as a rule to use sample estimates down to a population of approximately 200, instead of 500. The James-Stein procedures here, however, allow a combination of the two estimates to achieve an improvement in the average accuracy of prediction.

Table 1 also shows that regressions, involving either IRS or housing data, but especially those including both, are significantly more effective in estimating the true values than the regression on the county values alone. The fit for these other regressions is particularly good among states in the North Central Region. (One large value of A^* for Georgia is based on a relatively small number of cases.)

Before processing the entire set of estimates, we experimented with alternative forms for the regression equations, using the value of A^* as the criterion. Surprisingly, we did not find any appreciable improvement through further transformation of the independent variables.

Table 2 displays values of A^* obtained for places between 500 and 999. The values in the table tend to be somewhat less than those in the first table, indicating slightly better fit for larger places. The differences between Tables 1 and 2, however, are less than the difference between the average sampling errors of these two groups of places. Roughly speaking, places with less than 500 would have an average size of 250, while places between 500 and 999 would have an average size of about 750. Thus, the average sampling variances might differ by a factor of up to 3 between the two groups, while the ratios between the average estimated A^*'s are about 1.5. Thus, the assumption that the prior variance A^* is independent of D_i seems to hold reasonably, although not perfectly. Furthermore, possible inadequacies in the approximation used to give the sampling variances may affect the estimates in Table 1. In general, overestimates of the sampling variances will lead to underestimates of A^*.

For cases in which there may be some linkage between the sampling variance D_i and the variation of the true values about the predicted values, we include in the Appendix a procedure to fit the assumption $\theta_i \sim_{\text{ind}} N(X_i\beta,\ AD_i^a)$. Use of the procedure would be encouraged, however, only if many cases, perhaps on the order of hundreds, were available and the true values of D_i were known to almost complete accuracy.

4. EVALUATION OF THE ESTIMATOR

The values of A^* indicated that the revised estimator would be superior to the county values. In some applications, these statistics may constitute the only available assessment of the improvement achieved by the James-Stein estimator, where small values of A^* relative to the sampling variances D_i point to substantial overall gains. For this problem, however, we devised two additional demonstrations of characteristics of the revised estimator: one based on a limited number of special censuses taken in 1973, and the other derived from the 1970 data used in the estimation.

As a general verification of the methodology to update the 1970 census estimates of population and income on the basis of changes in administrative data, the Census Bureau conducted complete censuses of a random sample of places and townships in 1973, collecting income for 1972 on a 100 percent basis. (The difference in years here is the same as for the 1970 census collecting income for 1969; in general, Census Bureau income questions are asked for income during the preceding calendar year.) Of these special censuses, 17 were for places of size less than 500 in 1970, and 7 fell into the interval 500 to 999. In general, the methodology to update the estimates produced for each place a factor f_i used to multiply a base figure for 1969. By keeping this updating factor f_i constant, three separate estimates of PCI for 1972 were

3. Comparison of Selected 1972 PCI Estimates With 1973 Special Census Values of 1972 PCI

Special Census Areas	1973 Special Census 1972 PCI	1972 PCI Estimates and Percentage Difference From Special Census PCI: Using 1970 Sample Base — 1972 Estimate	Using 1970 Sample Base — Percentage Difference	Using Revised Base (James-Stein) — 1972 Estimate	Using Revised Base (James-Stein) — Percentage Difference	Using County Base — 1972 Estimate	Using County Base — Percentage Difference
1970 Census Weighted Sample Population Less Than 500							
Newington, Ga.	$2,019	$2,225	10.2	$2,302	14.0	$2,279	12.9
Foosland Village, Ill.	2,899	2,771	4.4	3,199	10.3	3,796	30.9
Bonaparte, Iowa	2,331	3,126	34.1	2,942	26.2	2,542	9.1
McNary, La.	2,333	2,303	1.3	2,527	8.3	2,908	24.6
Freeborn Village, Minn.	2,741	3,693	34.7	3,338	21.8	2,922	6.6
Spruce Valley Twp., Minn.	2,430	1,894	22.1	1,949	19.8	2,076	14.6
Jacksonville, Mo.	2,723	2,338	14.1	2,611	4.1	3,233	18.7
Thayer, Nebr.	2,742	2,245	18.1	2,870	4.7	3,452	25.9
Benton Town, N.H.	1,788	2,874	60.7	3,284	78.7	3,570	99.7
Nora Twp., N.Dak.	1,780	2,629	47.7	2,754	54.7	3,476	95.3
Riga Twp., N.Dak.	1,454	2,749	89.1	2,411	65.8	2,711	86.5
Deer Creek, Okla.	2,451	2,493	1.7	2,673	9.1	2,762	12.7
Dudley Borough, Pa.	2,446	2,168	11.4	2,411	1.4	2,608	6.6
Brookings Twp., S.Dak.	3,132	3,400	8.6	3,309	5.7	2,395	23.5
Valley Twp., S.Dak.	1,574	1,946	23.6	1,972	25.3	2,114	34.3
Bryant Twp., S.Dak.	2,412	1,120	53.6	2,158	10.5	2,695	11.7
Parrish Town, Wis.	3,567	5,399	51.4	4,079	14.4	2,721	23.7
Average Percentage Difference	—	—	28.6	—	22.0	—	31.6
1970 Census Weighted Sample Population Between 500 and 999							
Caswell Plantation, Maine	$1,946	$2,656	36.5	$2,490	28.0	$2,646	36.0
Sugar Creek Twp., Mo.	2,224	2,035	8.5	2,315	4.1	2,018	9.3
Jeromesville, Ohio	3,329	3,081	7.4	3,418	2.7	3,072	7.7
Rush Twp., Ohio	2,241	2,545	13.6	2,619	16.9	2,546	13.6
Dennison Twp., Pa.	3,521	4,411	25.3	4,095	16.3	4,430	25.8
Manor, Tex.	2,062	2,746	33.2	2,765	34.1	2,740	32.9
Derby Center, Vt.	2,968	2,694	9.2	2,754	7.2	2,675	9.9
Average Percentage Difference	—	—	19.1	—	15.6	—	19.3

possible: multiplying the census sample estimate by f_i; multiplying the revised James-Stein estimate for 1969 by f_i; or multiplying the county values by f_i. The last, of course, was the original choice for the 1972 Revenue Sharing estimates. Comparison of the three sets of 1972 estimates with the special census results provides an indirect assessment of three sets of estimates for 1969, because each set is affected by errors both in the bases and in the updating factors f_i. Table 3 presents the results. The revised James-Stein estimator shows smaller average errors and, to a lesser extent, a lower incidence of extreme error than either the sample estimates or the county values. (The reader may note, however, that the estimates, particularly for the revised James-Stein base, run consistently higher than the special census values. The explanation lies with the special censuses themselves. Approximately 60 additional special censuses not included in this table were taken at the same time for places with population greater than 1,000, where the 1970 census sample estimates are used as base figures. There too, the estimates fall slightly above the special census results. One factor possibly involved is that missing income was not imputed in the processing of the special censuses, while it was in the 1970 census. The special censuses estimates, which are based on only complete cases, may be subject to a downward bias for this reason.)

A second test illustrates the manner in which the revised estimates, far more than the county values,

4. Relation of 1969 Revised Estimates and 1969 County Averages to 1970 Census Sample Estimates for Groups of 10

Relation to 1969 Sample Estimates	1969 Revised Estimates — Number	1969 Revised Estimates — Percentage	1969 County Averages — Number	1969 County Averages — Percentage
Total Groups	212	100.0	212	100.0
Within 10% of Sample PCI	172	81.1	111	52.4
Outside 10% of Sample PCI	40	18.9	101	47.6
Within One Standard Error	149	70.3	61	28.8
Between One and Two Standard Errors	28	13.2	60	28.3
Outside Two Standard Errors	35	16.5	91	42.9
Closer to Sample PCI	154	72.6	58	27.4

NOTE: For places with the ratio of 1969 IRS exemptions to 1970 census population between .8 and 1.1.

preserve much of the underlying dispersion of the true values for PCI among places. The logic of the test was simple: Although the estimates for places with population less than 500 have large sampling errors in the 1970 census, if they were pooled into a suitably large group, the sampling error of the resulting estimated sum would be relatively less. We selected a sample of places with valid IRS estimates of adjusted gross income per exemption and assembled them into groups of 10 after sorting them by the IRS values. (This grouping is legitimate in the sense that the groups are defined independently of the sample selection in the census.) Table 4 shows that the sum of the revised James-Stein estimates for 1969 more often falls closer to the sum of the census estimates than does the sum of the county values. This demonstration illustrates how the James-Stein estimates capture more of the true differences in income among these places than does the substitution of county values.

5. DISCUSSION

The General Revenue Sharing Program is one of a number of important federal programs that allocates funds according to formulas using statistical counts or estimates. A general study of critical considerations in the design and administration of these programs has recently appeared, *Statistical Policy Working Paper 1: Report on Statistics for Allocation of Funds* (U.S. Department of Commerce 1978). Although this study principally concerns policy issues beyond the scope of this article, the report reaffirms the need for accurate and timely data in these allocation programs.

The use of a James-Stein estimator in this instance does not form the precedent for its wholesale application by the Census Bureau to all other estimation problems involving small-area data, and we should emphasize the special circumstances in this application. Planning for the 1970 Census of Population and Housing did not anticipate the requirements of the State and Local Fiscal Assistance Act of 1972, which mandated the General Revenue Sharing Program. Consequently, the legislation forced a provisional program of estimation falling short of the ideal in this case, a 100 percent census for small places. The requirements of the act are in consideration in the planning of the 1980 census. In the meantime, however, the lack of sample estimates with acceptable statistical reliability forces a choice among alternatives, and in this instance James-Stein procedures provide an attractive solution.

Future applications may be required under certain conditions. For example, if the 1980 census is conducted with a 50 percent sample in small places, the sampling reliability of the estimates will be much greater than in 1970. Nonetheless, the sampling errors for places of size 200 or less may suggest a repetition of the same methodology. Parallel situations may arise in which the sampling errors of census or survey estimates may require the consideration of alternate estimators. In these instances, the James-Stein procedures may be viewed as a way to maximize the use of the data rather than as a means to replace them.

Although the theory is now sufficient to form the basis for many applications of the James-Stein estimator to practical problems, this article illustrates directions for further research. The general problem of unequal variances will require more investigation to produce estimators with good properties for practical applications. The question of how independent estimation problems may be grouped for joint consideration was partially addressed by Efron and Morris (1973b), but we conjecture that unequal variances introduce further complications in this question and that larger groupings than Efron and Morris suggested may have merit. (In retrospect, we would now contemplate further dividing the estimation problem along the dimension of population size in states with large numbers of small places, but we hesitated doing this initially in an effort to keep the procedures as uniform across states as possible.) The use of prior distributions other than the normal to motivate an empirical Bayes estimator may produce somewhat better results for practical problems in which only a few of the observations lie far from the general tendencies. Full solution of these problems may encourage further application of these techniques to practical problems.

APPENDIX

A1. Iterative Solution of (2.19) and (2.21)

Equation (2.19) for any specific value of A is, of course, simply weighted linear regression. To denote the dependence of (2.21) on the value of A, we will write $Y_i^*(A)$. Using the functions

$$f(A_n) = \sum_i \frac{(Y_i - Y_i^*(A_n))^2}{A_n + D_i} \tag{A.1}$$

$$g(A_n) = -\sum_i \frac{(Y_i - Y_i^*(A_n))^2}{(A_n + D_i)^2} \tag{A.2}$$

we started with $A_0 = 0$ and defined $A_{n+1} = A_n + (k - p - f(A_n))/g(A)$, constraining $A_{n+1} \geq 0$. The function g is an approximation to the derivative of f. Convergence is rapid, generally requiring less than 10 steps.

A2. Alternate Estimators

Carl Morris suggested to us a maximum likelihood approach to estimating A, which requires the simultaneous solution of (2.19) and

$$\sum_i \frac{(Y_i - Y_i^*)^2}{(A + D_i)^2} = \sum_i \frac{1}{A + D_i}. \tag{A.3}$$

Equations (A.3) and (2.21) weight the significance of the deviations $(Y_i - Y_i)^2$ differently: (A.3) places relatively more importance on the observations with small D_i than does (2.21). The maximum likelihood approach improves the efficiency of the estimation of A in the full Bayes setting. We preferred in this application,

however, to balance out the estimation of A over the sample to ensure that A was representative of all places in the class less than 500 rather than of just the larger places.

A more complex model may be fitted for cases in which both the assumed Bayes variance of the population and the sampling variance are related to a measure of size. In such cases, $Y_i \sim_{\text{ind}} N(\theta_i, D_i)$ and $\theta_i \sim_{\text{ind}} N(\mathbf{X}_i\beta, AD_i^\alpha)$ may be a reasonable model. Maximum likelihood techniques may be used to estimate β, A, and α jointly, requiring solution of (2.19) and

$$\sum_i \frac{(Y_i - Y_i^*)^2 D_i^\alpha}{(D_i + AD_i^\alpha)^2} = \sum_i \frac{D_i^\alpha}{D_i + AD_i^\alpha} \qquad \text{(A.4)}$$

$$\sum_i \frac{(Y_i - Y_i^*)^2 D_i^\alpha \ln D_i}{(D_i + AD_i^\alpha)^2} = \sum_i \frac{D_i^\alpha \ln D_i}{(D_i + AD_i^\alpha)^2}. \qquad \text{(A.5)}$$

We would conjecture, however, that the sampling error of α is too large to make this estimator preferable to simpler versions unless many, possibly several hundred, observations were involved.

A3. Details of the Implementation

For each place, the data were edited according to the following rules:

1. The census sample estimates were considered to be missing if the sample estimate of the number of persons was zero, or if the estimated PCI was less than $200. The latter situation can arise from losses, particularly on farm income, but the difficulty of assigning a reasonable standard error to the estimate in this instance led us to exclude such cases.
2. The housing data were considered missing if more than 20 percent of the owner-occupied units were farm dwellings, or if the data were otherwise unavailable.
3. The IRS data (originally prepared according to 1972 geography) were considered missing if the boundary changes between 1969 and 1972 had involved more than a 10 percent change in population, or if the number of exemptions was less than 70 percent or more than 100 percent of the 100 percent census count.

The equations incorporating county, tax, and housing values were calculated only for states with 16 or more complete cases. Similarly, regressions with either tax or housing data only were fitted for 12 or more valid cases and the county-only regressions required at least 8.

[*Received December 1977. Revised December 1978.*]

REFERENCES

Carter, Grace M., and Rolph, John E. (1974), "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," *Journal of the American Statistical Association*, 69, 880–885.

Draper, N.R., and Smith H. (1966), *Applied Regression Analysis*, New York: John Wiley & Sons.

Efron, Bradley, and Morris, Carl (1971), "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part I: The Bayes Case," *Journal of the American Statistical Association*, 66, 807–815.

——— (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, 130–139.

——— (1973a), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

——— (1973b), "Combining Possibly Related Estimation Problems," *Journal of the Royal Statistical Society*, Ser. B, 35, 379–421.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Ericksen, Eugene P. (1973), "A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," *Demography*, 10, 137–160.

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

Herriot, Roger A. (1977), "Updating Per Capita Income for General Revenue Sharing," in *Small Area Statistics Papers, Series GE-41, No. 4*, U.S. Bureau of the Census.

James, W., and Stein, Charles (1961), "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1*, Berkeley: University of California Press, 361–379.

Madow, W.G., and Hansen, M.H. (1975), "On Statistical Models and Estimation in Sample Surveys," in *Contributed Papers, 40th Session of the International Statistical Institute*, Warsaw, Poland, 554–557.

Rao, C. Radhakrishna (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Stein, Charles (1955), "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, Berkeley: University of California Press, 197–202.

U.S. Bureau of the Census (1976), *U.S. Census of Population and Housing: 1970, Procedural History PHC(R)-1*, Washington, D.C.

U.S. Department of Commerce (1978), *Statistical Policy Working Paper 1: Report on Statistics for Allocation of Funds*, prepared by the Subcommittee on Statistics for Allocation of Funds, Federal Committee on Statistical Methodology.

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 67-72
Statistics Canada

On Robust Small Area Estimation Using a Simple Random Effects Model

N.G.N. PRASAD and J.N.K. RAO[1]

ABSTRACT

Robust small area estimation is studied under a simple random effects model consisting of a basic (or fixed effects) model and a linking model that treats the fixed effects as realizations of a random variable. Under this model a model-assisted estimator of a small area mean is obtained. This estimator depends on the survey weights and remains design-consistent. A model-based estimator of its mean squared error (MSE) is also obtained. Simulation results suggest that the proposed estimator and Kott's (1989) model-assisted estimator are equally efficient, and that the proposed MSE estimator is often much more stable than Kott's MSE estimator, even under moderate deviations of the linking model. The method is also extended to nested error regression models.

KEY WORDS: Design consistent; Linking model; Mean squared error; Survey weights.

1. INTRODUCTION

Unit-level random effects models are often used in small area estimation to obtain efficient model-based estimators of small area means. Such estimators typically do not make use of the survey weights (*e.g.*, Ghosh and Meeden 1986; Battese, Harter and Fuller 1988; Prasad and Rao 1990). As a result, the estimators are not design consistent unless the sampling design is self-weighting within areas. We refer the reader to Ghosh and Rao (1994) for an appraisal of small area estimation methods.

Kott (1989) advocated the use of design-consistent model-based estimators (*i.e.*, model assisted estimators) because such estimators provide protection against model failure as the small area sample size increases. He derived a design-consistent estimator of a small area mean under a simple random effects model. This model has two components: the basic (or fixed effects) model and the linking model. The basic model is given by

$$y_{ij} = \theta_i + e_{ij},\ j = 1, 2, ..., N_i;\ i = 1, 2, ..., m \qquad (1)$$

where the y_{ij} are the population values and the e_{ij} are uncorrelated random errors with mean zero and variance σ_i^2 for each small area $i (= 1, 2, ..., m)$. For simplicity, we take θ_i as the small area mean $\bar{Y}_i = \sum_j y_{ij} / N_i$, where N_i is the number of population units in the i-th area. Note that $\bar{Y}_i = \theta_i + \bar{E}_i$ and $\bar{E}_i = \sum_j e_{ij} / N_i \approx 0$ if N_i is large.

The linking model assumes that θ_i is a realization of a random variable satisfying the model

$$\theta_i = \mu + v_i \qquad (2)$$

where the v_i are uncorrelated random variables with mean zero and variance σ_v^2. Further, $\{v_i\}$ and $\{e_{ij}\}$ are assumed to be uncorrelated.

Assuming that the model (1) also holds for the sample $\{y_{ij}, j = 1, 2, ..., n_i;\ i = 1, 2, ..., m\}$ and combining the sample model with the linking model, Kott (1989) obtained the familiar unit-level random effects model

$$y_{ij} = \mu + v_i + e_{ij},\ j = 1, 2, ..., n_i;\ i = 1, 2, ..., m, \qquad (3)$$

also called the components-of-variance model. It is customary to assume equal variances $\sigma_i^2 = \sigma^2$, although the case of random error variances has also been studied (Kleffe and Rao 1992; Arora and Lahiri 1997).

Assuming $\sigma_i^2 = \sigma^2$, Kott (1989) derived an efficient estimator $\hat{\theta}_{iK}$ of θ_i which is both model-unbiased under (3) and design-consistent. He also proposed an estimator of its mean squared error (MSE) which is model unbiased under the basic model (1) as well as design-consistent. But this MSE estimator can be quite unstable and can even take negative values, as noted by Kott (1989) in his empirical example. Kott (1989) used his MSE estimators mainly to compare the overall reduction in MSE from using $\hat{\theta}_{iK}$ in place of a direct design-based estimator $\bar{y}_{iw}$ given by (4) below. He remarked that more stable MSE estimators are needed.

The main purpose of this paper is to obtain a pseudo empirical best linear unbiased prediction (EBLUP) estimator of θ_i which depends on the survey weights and is design-consistent (section 2). A stable model-based MSE estimator is also obtained (section 3). Results of a simulation study in section 4 show that the proposed MSE estimator is often much more stable than the MSE estimator of Kott, as measured by their coefficient of variation, even under moderate deviations of the linking model (2). Results under the simple model (3) are also extended to a nested error regression model (section 5).

[1] N.G.N. Prasad, Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1; J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

2. PSEUDO EBLUP ESTIMATOR

Suppose $\tilde{w}_{ij}$ denotes the basic design weight attached to the j-th sample unit ($j = 1, 2, ..., n_i$) in the i-th area ($i = 1, 2, ..., m$). A direct design-based estimator of θ_i is then given by the ratio estimator

$$\bar{y}_{iw} = \sum_j \tilde{w}_{ij} y_{ij} \Big/ \sum_j \tilde{w}_{ij} = \sum_j w_{ij} y_{ij} \tag{4}$$

where $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$. The direct estimator $\bar{y}_{iw}$ is design-consistent but fails to borrow strength from the other areas.

To get a more efficient estimator, we consider the following reduced model obtained from the combined model (3) with $\sigma_i^2 = \sigma^2$:

$$\begin{aligned} \bar{y}_{iw} &= \sum_j w_{ij}(\mu + v_i + e_{ij}) \\ &= \mu + v_i + \bar{e}_{iw}, \end{aligned} \tag{5}$$

where the $\bar{e}_{iw}$ are uncorrelated random variables with mean zero and variance $\delta_i = \sigma^2 \sum_j w_{ij}^2$. The reduced model (5) is an area-level model similar to the well-known Fay-Herriot model (Fay and Herriot 1979). It now follows from the standard best linear unbiased prediction (BLUP) theory (*e.g.*, Prasad and Rao 1990) that the BLUP estimator of $\theta_i = \mu + v_i$ for the reduced model (5) is given by

$$\tilde{\theta}_i = \tilde{\mu}_w + \tilde{v}_i, \tag{6}$$

where

$$\tilde{v}_i = \gamma_{iw}(\bar{y}_{iw} - \tilde{\mu}_w)$$

with $\tilde{\mu}_w = \sum_i \gamma_{iw} \bar{y}_{iw} / \sum_i \gamma_{iw}$ and $\gamma_{iw} = \sigma_v^2 / (\sigma_v^2 + \delta_i)$. Note that $\tilde{\theta}_i$ is different from the BLUP estimator under the full model (3). We therefore denote $\tilde{\theta}_i$ as a pseudo-BLUP estimator. The estimator (6) may also be written as a convex combination of the direct estimator $\bar{y}_{iw}$ and $\tilde{\mu}_w$:

$$\tilde{\theta}_i = \gamma_{iw} \bar{y}_{iw} + (1 - \gamma_{iw}) \tilde{\mu}_w. \tag{7}$$

The estimator $\tilde{\theta}_i$ depends on the parameters σ_v^2 and σ^2 which are generally unknown in practice. We therefore replace σ_v^2 and σ^2 in (7) by model-consistent estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ under the original unit-level model (3) to obtain the estimator

$$\hat{\theta}_i = \hat{\gamma}_{iw} \bar{y}_{iw} + (1 - \hat{\gamma}_{iw}) \hat{\mu}_w, \tag{8}$$

where

$$\hat{\gamma}_{iw} = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}^2 \sum_j w_{ij}^2 \right)$$

and

$$\hat{\mu}_w = \sum_i \hat{\gamma}_{iw} \bar{y}_{iw} \Big/ \sum_i \hat{\gamma}_{iw}.$$

The estimator $\hat{\theta}_i$ will be referred to as pseudo-EBLUP estimator. We use standard estimators of σ_v^2 and σ^2, based on the within-area sums of squares

$$Q_w = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

and the between-area sums of squares

$$Q_b = \sum_i n_i (\bar{y}_i - \bar{y})^2,$$

where $\bar{y} = \sum_i n_i \bar{y}_i / \sum_i n_i$ is the overall sample mean. We have

$$\hat{\sigma}^2 = Q_w \Big/ \left(\sum_i n_i - m \right)$$

and $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ where

$$\tilde{\sigma}_v^2 = [Q_b - (m - 1)\hat{\sigma}^2] / n^*$$

with

$$n^* = \sum_i n_i - \sum_i n_i^2 \Big/ \sum_i n_i.$$

It may be noted that σ_v^2 and σ^2 are either not estimable or poorly estimated from the reduced model (5) due to identifiability problems. Following Kackar and Harville (1984), it can be shown that the pseudo-EBLUP estimator $\hat{\theta}_i$ is model-unbiased for θ_i under the original model (3) for symmetrically distributed errors $\{v_i\}$ and $\{e_{ij}\}$, not necessarily normal. It is also design consistent, assuming that $n_i \sum_j w_{ij}^2$ is bounded as n_i increases, because $\hat{\gamma}_{iw}$ converges in probability to 1 as $n_i \to \infty$ regardless of the validity of the model (3), assuming $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ converge in probability to some values, say, $\hat{\sigma}_v^{*2}$ and σ^{*2}.

Kott's (1989) model-based estimator of θ_i is obtained by taking a weighted combination of $\bar{y}_{iw}$ and $\sum_{l \neq i} c_l^{(i)} \bar{y}_l$, that is,

$$f_i(\alpha_i, c^{(i)}) = (1 - \alpha_i) \bar{y}_{iw} + \alpha_i \sum_{l \neq i} c_l^{(i)} \bar{y}_l,$$

and then minimizing the model mean squared error (MSE) of $f_i(\alpha_i, c^{(i)})$ with respect to α_i and $c_l^{(i)}$ subject to model-unbiasedness condition: $\sum_{l \neq i} c_l^{(i)} = 1$. This leads to

$$\hat{\theta}_{iK} = f_i(\hat{\alpha}_i, \hat{c}^{(i)}) \tag{9}$$

with

$$\hat{\alpha}_i = \sum_j \left[w_{ij}^2 \Big/ \left\{ \sum_j w_{ij}^2 + \sum_{l \neq i} \hat{c}_l^{(i)2} \Big/ n_i + \left(1 + \sum_{l \neq i} \hat{c}_l^{(i)2} \right) \left(\hat{\sigma}_v^2 / \hat{\sigma}^2 \right) \right\} \right]$$

and

$$\hat{c}_l^{(i)} = \left[(\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_l^{-1} \right] \Big/ \sum_{h \neq i} \left[(\hat{\sigma}_v^2 / \hat{\sigma}^2) + n_h^{-1} \right].$$

The estimator $\hat{\theta}_{iK}$ is also model-unbiased and design-consistent. In a previous version of this paper, we proposed an estimator similar to (9). It uses the best estimators of μ under the unit-level model, based on the unweighted means $\bar{y}_i$, rather than $\hat{\mu}_w$, the best estimator of μ under the reduced model (4), based on the survey-weighted means $\bar{y}_{iw}$.

3. ESTIMATORS OF MSE

It is straightforward to derive the MSE of the pseudo-BLUP estimator $\tilde{\theta}_i$ under the unit level model (3). We have

$$\text{MSE}(\tilde{\theta}_i) = E(\tilde{\theta}_i - \theta_i)^2 = g_{1i}(\sigma_v^2, \sigma^2) + g_{2i}(\sigma_v^2, \sigma^2) \quad (10)$$

with

$$g_{1i}(\sigma_v^2, \sigma^2) = (1 - \gamma_{iw})\sigma_v^2$$

and

$$g_{2i}(\sigma_v^2, \sigma^2) = \sigma_v^2(1 - \gamma_{iw})^2 \Big/ \sum_i \gamma_{iw}.$$

The leading term, $g_{1i}(\sigma_v^2, \sigma^2)$ is of order $O(1)$, while the second term, $g_{2i}(\sigma_v^2, \sigma^2)$, due to estimation of μ is of order $O(m^{-1})$ for large m.

A naive MSE estimator of the pseudo-EBLUP estimator $\hat{\theta}_i$ is obtained by estimating $\text{MSE}(\hat{\theta}_i)$ given by (10):

$$\text{mse}_N(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \quad (11)$$

But (11) could lead to significant underestimation of $\text{MSE}(\hat{\theta}_i)$ because it ignores the uncertainty associated with $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$. Note that

$$\text{MSE}(\hat{\theta}_i) = \text{MSE}(\tilde{\theta}_i) + E(\hat{\theta}_i - \tilde{\theta}_i)^2 \quad (12)$$

under normality of the errors $\{v_i\}$ and $\{e_{ij}\}$ so that $\text{MSE}(\tilde{\theta}_i)$ is always smaller than $\text{MSE}(\hat{\theta}_i)$; see Kackar and Harville (1984).

To get a "correct" estimator of $\text{MSE}(\hat{\theta}_i)$, we first approximate the second order term $E(\hat{\theta}_i - \tilde{\theta}_i)^2$ in (12) for large m, assuming that $\{v_i\}$ and $\{e_{ij}\}$ are normally distributed. Following Prasad and Rao (1990), we have

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 \approx g_{3i}(\sigma_v^2, \sigma^2) \quad (13)$$

where the neglected terms are of lower order than m^{-1}, and

$$g_{3i}(\sigma_v^2, \sigma^2) = \gamma_{iw}(1 - \gamma_{iw})^2 \sigma_v^{-2} \{V(\tilde{\sigma}_v^2) - 2(\sigma_v^2/\sigma^2)\text{Cov}(\tilde{\sigma}_v^2, \hat{\sigma}^2) + (\sigma_v^2/\sigma^2)^2 \text{Var}(\hat{\sigma}^2)\}; \quad (14)$$

see Appendix 1. The variances and covariances of $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ are also given in the Appendix 1. It can be shown that $g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ is approximately unbiased for $g_{1i}(\sigma_v^2, \sigma^2)$ in the sense that its bias is of lower order than m^{-1} (see Appendix 2). Similarly, $g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ and $g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ are approximately unbiased for $g_{2i}(\sigma_v^2, \sigma^2)$ and $g_{3i}(\sigma_v^2, \sigma^2)$, respectively. It now follows that an approximately model-unbiased estimator of $\text{MSE}(\hat{\theta}_i)$ is given by

$$\text{mse}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2). \quad (15)$$

For the estimator $\hat{\theta}_{iK}$ given by (9), Kott (1989) proposed an estimator of MSE as

$$\text{mse}(\hat{\theta}_{iK}) = (1 - 2\hat{\alpha}_i)v^*(\bar{y}_{iw}) + \hat{\alpha}_i^2\left(\bar{y}_{iw} - \sum_{l \neq i} c_l^{(i)}\bar{y}_l\right)^2, \quad (16)$$

where $v^*(\bar{y}_{iw})$ is both a design-consistent estimator of the design-MSE of $\bar{y}_{iw}$ and a model-unbiased estimator of the model-variance of $\bar{y}_{iw}$ under the basic model (1). Since $\hat{\alpha}_i$ converges in probability to zero as $n_i \to \infty$, it follows from (16) that $\text{mse}(\hat{\theta}_{iK})$ is also both design-consistent and model unbiased assuming only the basic model (1). However, $\text{mse}(\hat{\theta}_{iK})$ is unstable and can even take negative values when $\hat{\alpha}_i$ exceeds 0.5, as noted by Kott (1989).

Note that our MSE estimator, $\text{mse}(\hat{\theta}_i)$ is based on the full model (3) obtained by combining the basic model (1) with the linking model (2). However, our simulation results in section 4 show that it may perform well even under moderate deviations from the linking model.

4. SIMULATION STUDY

We conducted a limited simulation study to evaluate the performances of the proposed estimator $\hat{\theta}_i$, given by (8), and its estimator of MSE, given by (15), relative to Kott's estimator $\hat{\theta}_{iK}$, given by (9), and its estimator of MSE, given by (16). We studied the performances under two different approaches: (i) For each simulation run, a finite population of $m = 30$ small areas with $N_i = 200$ population units in each area is generated from the assumed unit-level model and then a PPS (probability proportional to size) sample within each small area is drawn independently, using $n_i = 20$. (ii) A fixed finite population is first generated from the assumed unit-level model and then for each simulation run a PPS sample within each small area is drawn independently, employing the fixed finite population. Approach (i) refers to both the design and the linking model whereas approach (ii) is design-based in the sense that it refers only to the design. The errors $\{v_i\}$ and $\{e_{ij}\}$ are assumed to be normally distributed in generating the finite populations $\{y_{ij}, i = 1, 2, ..., 30; j = 1, 2, ..., 200\}$. We considered two cases: (1) The linking model (2) is true with $\mu = 50$. (2) The linking model is violated by letting μ vary across areas: $\mu_i = 50$, $i = 1, 2, ..., 10$; $\mu_i = 55$, $i = 11, 12, ..., 20$; $\mu_i = 60$,

$i = 21, 22, ..., 30$. To implement PPS sampling within each area, size measures $z_{ij} (i = 1, 2, ..., 30; j = 1, 2, ..., 200)$ were generated from an exponential distribution with mean 200. Using these z-values, we computed selection probabilities $p_{ij} = z_{ij} / \sum_j z_{ij}$ for each area i and then used them to select PPS with replacement samples of sizes $n_i = n$, by taking $n = 20$, and the associated sample values $\{y_{ij}\}$ were observed.

The basic design weights are given by $\tilde{w}_{ij} = n^{-1} p_{ij}^{-1}$ so that $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$. Using these weights and the associated sample values y_{ij} we computed estimates $\hat{\theta}_i$ and $\hat{\theta}_{iK}$ and associated estimates of MSE, and also the ratio estimate $\bar{y}_{iw}$ for each simulation run; the formula for $v^*(\bar{y}_{iw})$ under PPS sampling is given in Appendix 3. This process was repeated $R = 10{,}000$ times to get from each run $r (= 1, 2, ..., R)$ $\hat{\theta}_i(r)$ and $\hat{\theta}_{iK}(r)$ and associated MSE estimates $\text{mse}_i(\hat{\theta}_i(r))$ and $\text{mse}_i(\hat{\theta}_{iK}(r))$ and also the direct estimate $\bar{y}_{iw}(r)$. Using these values, empirical relative efficiencies (RE) of $\hat{\theta}_i$ and $\hat{\theta}_{iK}$ over $\bar{y}_{iw}$ were computed as

$$\text{RE}(\hat{\theta}_i) = \text{MSE}_*(\bar{y}_{iw}) / \text{MSE}_*(\hat{\theta}_i)$$

and

$$\text{RE}(\hat{\theta}_{iK}) = \text{MSE}_*(\bar{y}_{iw}) / \text{MSE}_*(\hat{\theta}_{iK}),$$

where MSE_* denotes the MSE over $R = 10{,}000$ runs. For example, $\text{MSE}_*(\hat{\theta}_i) = \sum_r [\hat{\theta}_i(r) - \bar{Y}_i(r)]^2 / R$, where $\bar{Y}_i(r)$ is the i-th area population mean for the r-th run. Note that $\bar{Y}_i(r)$ remains the same over the runs r under the design-based approach because the finite population is fixed over the simulation runs.

Similarly, the relative biases of the MSE estimators were computed as

$$\text{RB}[\text{mse}(\hat{\theta}_i)] = [\text{MSE}_*(\hat{\theta}_i) - E_*\text{mse}(\hat{\theta}_i)] / \text{MSE}_*(\hat{\theta}_i)$$

and

$$\text{RB}[\text{mse}(\hat{\theta}_{iK})] = [\text{MSE}_*(\hat{\theta}_{iK}) - E_*\text{mse}(\hat{\theta}_{iK})] / \text{MSE}_*(\hat{\theta}_{iK}),$$

where E_* denotes the expectation over $R = 10{,}000$ runs. For example, $E_*\text{mse}(\hat{\theta}_i) = \sum_r \text{mse}(\hat{\theta}_i(r)) / R$. Finally, the empirical coefficient of variation (CV) of the MSE estimators were computed as

$$\text{CV}[\text{mse}(\hat{\theta}_i)] = [\text{MSE}_*\{\text{mse}(\hat{\theta}_i)\}]^{1/2} / \text{MSE}_*(\hat{\theta}_i)$$

and

$$\text{CV}[\text{mse}(\hat{\theta}_{iK})] = [\text{MSE}_*\{\text{mse}(\hat{\theta}_{iK})\}]^{1/2} / \text{MSE}_*(\hat{\theta}_{iK}).$$

Note that $\text{MSE}_*[\text{mse}(\hat{\theta}_i)] = \sum_r [\text{mse}(\hat{\theta}_i(r)) - \text{MSE}_*(\hat{\theta}_i)]^2 / R$ and a similar expression for $\text{MSE}_*[\text{mse}(\hat{\theta}_{iK})]$.

Table 1 reports summary measures of the values of percent RE, |RB| and CV for cases (1) and (2) under approach (i). Summary measures under approach (ii) are reported in Table 2. Summary measures considered are the mean and the median (med) over the small areas $i = 1, 2, ..., 30$.

Table 1
Relative Efficiency (RE) of Estimators, Absolute Relative Bias (|RB|) and Coefficient of Variation (CV) of MSE estimators (σ=5.0, n=20): Approach (i)

σ_v		RE%		|RB|%		CV%	
		$\hat{\theta}_{iK}$	$\hat{\theta}_i$	$\text{mse}(\hat{\theta}_{iK})$	$\text{mse}(\hat{\theta}_i)$	$\text{mse}(\hat{\theta}_{iK})$	$\text{mse}(\hat{\theta}_i)$
				Case 1			
1	Mean	190	177	15.3	3.5	148	25
	Med	190	182	14.8	2.6	148	25
2	Mean	126	123	5.1	3.2	48	8
	Med	127	124	5.6	2.9	48	8
3	Mean	113	111	3.5	2.7	35	6
	Med	112	111	3.2	3.0	35	6
				Case 2			
1	Mean	108	103	10.4	7.9	39	6
	Med	108	104	11.1	7.7	38	5
2	Mean	108	104	13.3	8.9	39	6
	Med	108	104	13.6	7.9	37	6
3	Mean	104	103	11.5	7.2	37	5
	Med	105	105	13.1	8.0	36	6

Case 1: μ_i=50, i=1,2,...,30; Case 2: μ_i=50, i=1,2,...,10; μ_i=55, i=11,12,...,20; μ_i=60, i=21,22,...,30.

It is clear from Tables 1 and 2 that $\hat{\theta}_{iK}$ and $\hat{\theta}_i$ perform similarly with respect to RE which decreases as σ_v / σ increases. Under approach (ii), RE is large for both cases 1 and 2 when $\sigma_v / \sigma \leq 0.4$, whereas it decreases significantly under approach (i) if the linking model is violated (case 2); the direct estimator $\bar{y}_{iw}$ is quite unstable under approach (ii).

Turning to the performance of MSE estimators under approach (i), Table 1 shows that |RB| of $\text{mse}(\hat{\theta}_i)$ is negligible (<4%) when the linking model holds (Case 1) and that it is small (<10%) even when the linking model is violated, although it increases. The estimator $\text{mse}(\hat{\theta}_{iK})$ has a larger |RB| but it is less than 15%. The CV of $\text{mse}(\hat{\theta}_i)$ is much smaller than the CV of $\text{mse}(\hat{\theta}_{iK})$ for both Cases 1 and 2. For example, when the model holds (Case 1) the median CV is 25% for $\text{mse}(\hat{\theta}_i)$ compared to 148% for $\text{mse}(\hat{\theta}_{iK})$ when $\sigma_v = 1$; the median CV decreases to 8% for $\text{mse}(\hat{\theta}_i)$ compared to 48% for $\text{mse}(\hat{\theta}_{iK})$ when $\sigma_v = 2$. This pattern is retained when the model is violated (Case 2). It may be noted that the probability of $\text{mse}(\hat{\theta}_{iK})$ taking a negative value is quite large (>0.3) when $\sigma_v / \sigma \leq 0.4$.

Under approach (ii), Table 2 shows that |RB| of $\text{mse}(\hat{\theta}_i)$ is larger than the value under approach (i) and ranges from 15% to 25%. On the other hand, |RB| of $\text{mse}(\hat{\theta}_{iK})$ is smaller and ranges from 4% to 15%. The CV of $\text{mse}(\hat{\theta}_{iK})$, how-

ever, is much larger than under approach (i). For example, the median CV for Case 1 is 295% compared to 38% for $\text{mse}(\hat{\theta}_i)$ when $\sigma_v = 1$ which decreases to 122% compared to 23% when $\sigma_v = 2$. A similar pattern holds for case 2 where the fixed finite population is generated from the model with varying means.

Table 2
Relative Efficiency (RE) of Estimators, Absolute Relative Bias (IRBI) and Coefficient of Variation (CV) of MSE estimators (σ=5.0, n=20): Approach (ii)

σ_v		RE%		IRBI%		CV%	
		$\hat{\theta}_{iK}$	$\hat{\theta}_i$	$\text{mse}(\hat{\theta}_{ik})$	$\text{mse}(\hat{\theta}_i)$	$\text{mse}(\hat{\theta}_{iK})$	$\text{mse}(\hat{\theta}_i)$
				Case 1			
1	Mean	283 281	14.2	25.4	289	39	
	Med	275 279	15.0	24.7	295	38	
2	Mean	180 182	7.3	19.2	115	24	
	Med	177 181	6.9	18.7	122	23	
3	Mean	129 129	4.8	14.8	68	24	
	Med	129 128	4.2	13.9	65	24	
				Case 2			
1	Mean	278 276	15.7	26.8	291	41	
	Med	271 275	16.6	26.2	297	40	
2	Mean	175 177	8.8	20.7	117	26	
	Med	173 177	8.5	20.3	124	25	
3	Mean	124 124	6.3	16.2	70	25	
	Med	125 124	6.8	15.5	67	26	

Case 1: μ_i=50, i=1,2,...,30; Case 2: μ_i=50, i=1,2,...,10; μ_i=55, i=11,12,...,20; μ_i =60, i=21,22,...,30.

To reduce IRBI of $\text{mse}(\hat{\theta}_i)$ under approach (ii), one could combine it with $\text{mse}(\hat{\theta}_{iK})$ by taking a weighted average, but it appears difficult to chose the appropriate weights. The weighted average will be more stable than $\text{mse}(\hat{\theta}_{iK})$.

5. NESTED ERROR REGRESSION MODEL

The results in sections 2 and 3 can be extended to nested error regression models

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij},\ j = 1, 2, ..., n_i;\ i = 1, 2, ..., m \qquad (17)$$

using the results of Prasad and Rao (1990), where x_{ij} is a p-vector of auxiliary variables with known population mean $\bar{X}_i$ and related to y_{ij}, and β is the p-vector of regression coefficients. The reduced model is given by

$$\bar{y}_{iw} = \bar{x}_{iw}'\beta + v_i + \bar{e}_{iw} \qquad (18)$$

with $\bar{x}_{iw}' = \sum_j w_{ij}x_{ij}$. Model-consistent estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ are obtained from the unit-level model (17), employing either the method of fitting constants (Prasad and Rao 1990) or REML (restricted maximum likelihood) estimation (Datta and Lahiri 1997).

The pseudo-EBLUP of $\theta_i = \bar{X}_i'\beta + v_i$ is given by

$$\hat{\theta}_i = \hat{\gamma}_{iw}\bar{y}_{iw} + (1 - \hat{\gamma}_{iw})\bar{X}_i'\hat{\beta}_w, \qquad (19)$$

where

$$\hat{\beta}_w = \left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}_{iw}'\right)^{-1}\left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{y}_{iw}\right).$$

An approximate model-unbiased estimator of $\text{MSE}(\hat{\theta}_i)$ is given by (15) with

$$g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) = (1 - \hat{\gamma}_{iw})\hat{\sigma}_v^2$$

as before,

$$g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) = \hat{\sigma}_v^2(\bar{X}_i - \hat{\gamma}_{iw}\bar{x}_{iw})'\left(\sum_i \hat{\gamma}_{iw}\bar{x}_{iw}\bar{x}_{iw}'\right)^{-1}(\bar{X}_i - \hat{\gamma}_{iw}x_{iw}\bar{x}_{iw})$$

and $g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$, obtained from (14), involves the estimated variances and covariances of $\tilde{\sigma}_v^2$ and $\hat{\sigma}^2$. The latter can be obtained from Prasad and Rao (1990) for the method of fitting constants and from Datta and Lahiri (1997) for REML.

6. CONCLUSION

We have proposed a model-assisted estimator of a small area mean under a simple unit-level random effects model. This estimator depends on the survey weights and is design-consistent. We have also obtained a model-based MSE estimator. Results of our simulation study have shown that the proposed MSE estimator performs well, even under moderate deviations of the linking model. The proposed approach is also extended to a nested error regression model.

ACKNOWLEDGEMENTS

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada. We are thankful to the Associate Editor and the referee for constructive comments and suggestions.

APPENDIX 1

Proof of (13):

From general results (Prasad and Rao 1990) we have

$$E(\hat{\theta}_i - \tilde{\theta}_i)^2 \approx tr\left[A_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)\right],$$

where $B_i(\sigma_v^2, \sigma^2)$ is the 2×2 covariance matrix of $\tilde{\sigma}_v^2$ and $\hat{\sigma}^2$, and $A_i(\sigma_v^2, \sigma^2)$ is the 2×2 covariance matrix of

$$\left(\frac{\partial\theta_i^*}{\partial\sigma_v^2}, \frac{\partial\theta_i^*}{\partial\sigma^2}\right).$$

Now, noting that

$$\frac{\partial\theta_i^*}{\partial\sigma_v^2} = \frac{\partial\gamma_{iw}}{\partial\sigma_v^2}\bar{y}_{iw} = \left[\frac{\gamma_{iw}(1-\gamma_{iw})}{\sigma_v^2}\right]\bar{y}_{iw},$$

$$\frac{\partial\theta_i^*}{\partial\sigma^2} = \frac{\partial\gamma_{iw}}{\partial\sigma^2}\bar{y}_{iw} = -\left[\frac{\gamma_{iw}(1-\gamma_{iw})}{\sigma^2}\right]\bar{y}_{iw},$$

and $V(\bar{y}_{iw}) = \sigma_v^2 + \delta_i = \sigma_v^2/\gamma_{iw}$, we get

$$A_i(\sigma_v^2, \sigma^2) = \left[\gamma_{iw}(1-\gamma_{iw})^2\sigma_v^{-2}\right]\begin{bmatrix} 1 & -\sigma_v^2/\sigma^2 \\ -\sigma_v^2/\sigma^2 & (\sigma_v^2/\sigma)^2 \end{bmatrix},$$

and hence the result (14).

Covariance matrix of $\tilde{\sigma}_v^2$ and $\hat{\sigma}^2$:

Under normality, we have

$$V(\hat{\sigma}^2) = 2\sigma^4/\left(\sum_i n_i - m\right),$$

$$V(\tilde{\sigma}_v^2) = 2n_{\cdot\cdot}^{-2}$$

$$\left[\sigma^4(m-1)\left(\sum n_i - 1\right)\left(\sum n_i - m\right)^{-1} + 2n_{\cdot\cdot}\sigma^2\sigma_v^2 + n_{\cdot\cdot}\sigma_v^4\right]$$

and

$$\mathrm{Cov}(\hat{\sigma}^2, \tilde{\sigma}_v^2) = -(m-1)n_{\cdot\cdot}^{-1}V(\hat{\sigma}^2),$$

where

$$n_{\cdot\cdot} = \sum n_i^2 - 2\sum n_i^3 \Big/ \sum n_i + \left(\sum n_i^2\right)^2 \Big/ \left(\sum n_i\right)^2;$$

see Searle, Casella and McCulloch (1992, p. 428).

APPENDIX 2

Proof of $E[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{1i}(\sigma_v^2, \sigma^2)$:

By a Taylor expansion of $g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ around (σ_v^2, σ^2) to second order and noting that $E(\hat{\sigma}^2 - \sigma^2) = 0$ and $E(\hat{\sigma}_v^2 - \sigma_v^2) \approx 0$, we get

$$E\left[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) - g_{1i}(\sigma_v^2, \sigma^2)\right] \approx \frac{1}{2}tr\left[D_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)\right],$$

where $D_i(\sigma_v^2, \sigma^2)$ is the 2×2 matrix of second order derivatives of $g_{1i}(\sigma_v^2, \sigma^2)$ with respect to σ_v^2 and σ^2. It is easy to verify that

$$\frac{1}{2}tr\left[D_i(\sigma_v^2, \sigma^2)B_i(\sigma_v^2, \sigma^2)\right] = g_{3i}(\sigma_v^2, \sigma^2).$$

Now, noting that $E[g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)] \approx g_{3i}(\sigma_v^2, \sigma^2)$ we get the desired result.

APPENDIX 3

The design-based estimator of variance of $\bar{y}_{iw}$ under PPS sampling is given by

$$v(\bar{y}_{iw}) = \frac{m}{m-1}\sum_j w_{ij}^2(y_{ij} - \bar{y}_{iw})^2.$$

Kott (1989) model-assisted variance estimator is

$$v^*(\bar{y}_{iw}) = \left\{V(\bar{y}_{iw})/Ev(\bar{y}_{iw})\right\}v(\bar{y}_{iw}) = \frac{\left(\sum_j w_{ij}^2\right)\sum_j w_{ij}^2(y_{ij} - \bar{y}_{iw})^2}{\sum_j w_{ij}^2\left(1 - 2w_{ij} + \sum_j w_{ij}^2\right)},$$

where E and V denote expectation and variance with respect to the basic model (1).

REFERENCES

ARORA,V., and LAHIRI, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

BATTESE, G.E., HARTER, R., and FULLER,W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

DATTA, G.S., and LAHIRI, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Technical Report, University of Nebraska-Lincoln.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects models. *Journal of the American Statistical Association*, 79, 853-862.

KLEFFE, J., and RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

KOTT, P. (1989). Robust small domain estimation using random effects modelling. *Survey Methodology*, 15, 3-12.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

SEARLE, S.R., CASELLA, G., and McCULLOCH, C.E. (1992). *Variance Components*. New York: John Wiley and Sons.

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 73-80
Statistics Canada

Small Area Estimation Using Multilevel Models

FERNANDO A.S. MOURA and DAVID HOLT[1]

ABSTRACT

In this paper a general multilevel model framework is used to provide estimates for small areas using survey data. This class of models allows for variation between areas because of: (i) differences in the distributions of unit level variables between areas, (ii) differences in the distribution of area level variables between areas and (iii) area specific components of variance which make provision for additional local variation which cannot be explained by unit-level or area-level covariates. Small area estimators are derived for this multilevel model formulation and an approximation to the mean square error (MSE) of each small area estimate for this general class of mixed models is provided together with an estimator of this MSE. Both the approximations to the MSE and the estimator of MSE take into account three sources of variation: (i) the prediction MSE assuming that both the fixed and components of variance terms in the multilevel model are known, (ii) the additional component due to the fact that the fixed coefficients must be estimated, and (iii) the further component due to the fact that the components of variance in the model must be estimated. The proposed methods are estimated using a large data set as a basis for numerical investigation. The results confirm that the extra components of variance contained in multilevel models as well as small area covariates can improve small area estimates and that the MSE approximation and estimator are satisfactory.

KEY WORDS: Small area estimation; Mixed models; Multilevel models; EBLUE.

1. INTRODUCTION

The need for small area (and small domain) estimates from survey data has long been recognized. The difficulty with the production of such estimates is that for most, if not all, small areas, the sample size achieved by a survey designed for national purposes is too small for direct estimates to be made with acceptable precision. Early attempts to tackle this problem using methods such as synthetic estimation (Gonzalez 1973) involved the use of auxiliary information and the pooling of information across small areas. An excellent review and bibliography are given by Ghosh and Rao (1994).

Empirical studies show that such methods made too little provision for local variation and consequently the resulting small area estimates were shrunk too far towards a predicted mean. More recent approaches (*e.g.*, Battese and Fuller 1981 and Battese, Harter and Fuller 1988) use some components of variance model, or equivalent, to provide for local variation. Empirical studies show the superiority of this approach (*e.g.*, Prasad and Rao 1990).

This paper proposes a general multilevel model framework for small area estimation. This involves the potential to use auxiliary information at both the unit and small area level. In addition any of the regression parameters, rather than just the intercept as proposed by Battese and Fuller (1981), may be treated as varying randomly between small areas. The local variation is provided for by using differences between the means of unit level auxiliary variables, the small area level variables, and the various components of variance which allow variation between areas.

For this general model, the small area predictor is obtained. In addition, an approximation to the mean square error (MSE) of each separate small area prediction and an estimator of this MSE are developed.

The numerical study, based on a large data set from Brazil shows that such models may be useful for predicting small area estimates. The robustness of the approach to misspecification of the variance-covariance matrix of the small area random effects and misspecification of small area covariates are also investigated. Further numerical results demonstrate the success of the MSE approximation and its estimator.

2. THE MULTILEVEL MODEL FRAMEWORK

2.1 Introduction

We consider the following multilevel model for predicting the small area means:

$$Y_i = X_i\beta_i + \varepsilon_i$$

$$\beta_i = Z_i\gamma + v_i \quad i = 1,...,m \tag{2.1}$$

where Y_i is the vector of length n_i for the characteristic of interest for the sample units in the i-th small area, $i = 1, ..., m$; X_i is the matrix of explanatory variables at sample unit level; Z_i is the design matrix of small area variables; γ is the vector of length q of fixed coefficients and $v_i = (v_{i0},...,v_{ip})^T$ is the vector of length $(p + 1)$ of random effects for the i-th small area. We assume the

[1] Fernando A.S. Moura, Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil, CP: 68530, CEP: 21941-590, e-mail: fmoura@dme.ufrj.br; David Holt, Office for National Statistics, 1 Drummond Gate, London, SW1P 2QQ, e-mail: tholt@ons.go.uk.

following about the distribution of the random vectors: (a) the v_i are independent between small areas and have a joint distribution within each small area with $E(v_i) = 0$ and (b) $V(v_i) = \Omega$ (b) The $\varepsilon_i s$ and $v_i s$ are independent and $V(\varepsilon_i) = \sigma^2 I$.

For the whole population (2.1) applies with n_i replaced by N_i, the small area population sizes.

The set of m equations in (2.1) can be concisely written by stacking them as

$$Y = XZ_\gamma + Xv + \varepsilon. \tag{2.2}$$

It is worth noting that the random intercept model (see section 2.3) can be regarded as a special case of the model (2.1) where Z_i is equal to the identity matrix for each small area and Ω has all terms constrained to be zero except the one corresponding to the variance of the intercept term. Other intermediate models exist, for instance, when Ω is diagonal so that the small area regression coefficients are random but uncorrelated between covariates.

Holt (in Ghosh 1994, page 82) observes that the advantage of the model (2.1) over other competitors is that it effectively integrates the use of unit level and area level covariates into a single model. Besides the use of extra random effects for the regression coefficients gives greater flexibility in situations where it is not appropriate to assume the same slope coefficients apply for all small areas.

2.2 Fixed and Component of Variance Parameter Estimates

The fixed and components of variance parameters in the model (2.1) are γ and $\theta = ([\text{Vech}(\Omega)]^T, \sigma^2)^T$ respectively. Various methods for estimating these model parameters in the case of a general mixed linear model are available. Most of them, based on iterative algorithms, lead to the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (RMLE) under certain regularity conditions.

Goldstein (1986) shows how consistent estimators can be obtained by applying iterative generalised least squares procedures (IGLS). He also proved its equivalence to the maximum likelihood estimator under normality. Later Goldstein (1989) proposed a slight modification of his algorithm (namely, restricted iterative generalised least squares (RIGLS)) which is equivalent to RMLE under normality. Unlike the IGLS estimates, the RIGLS estimation procedures provide unbiased estimates of the component of variance parameters by taking into account the loss in degrees of freedom resulting from estimating the fixed parameters.

This work is confined to the RIGLS approach as in Goldstein (1989). The RIGLS procedure is described in details in Appendix A.

2.3 The Estimator of the Small Area Mean

Assuming the model (2.1) and considering that the population size N_i in the i-th small area is large, we can write the mean for the i-th small area as

$$\mu_i = \bar{X}_i^T Z_i \gamma + \bar{X}_i^T v_i \tag{2.3}$$

where $\bar{X}_i$ is the $(p+1)$ population mean vector for the i-th small area.

An estimator of μ_i may be obtained by plugging the RIGLS estimators of γ and θ in the respective terms of equation (2.3), where the predictor of the i-th small area random effect v_i is given by $\hat{v}_i = \hat{\Omega} X_i^T \hat{V}_i^{-1} (Y_i - X_i Z_i \hat{\gamma})$ where $\hat{V}_i^{-1} = \hat{\sigma}^{-2} I - \hat{\sigma}^{-4} X_i \hat{\Omega} \hat{G}_i^{-1} X_i^T$ and $\hat{G}_i^{-1} = (I + \hat{\sigma}^{-2} X_i^T X_i \hat{\Omega})^{-1}$.

This estimator of μ_i is known as Empirical Best Linear Unbiased Estimator (EBLUE)

$$\hat{\mu}_i = \bar{X}_i^T Z_i \hat{\gamma} + \bar{X}_i^T \hat{v}_i. \tag{2.4}$$

Battese *et al.*, (1981, 1988) propose and apply a random intercept model to provide small area estimates. In this case, the Empirical Best Linear Unbiased Estimator is

$$\hat{\mu}_{i(\text{RI})} = \bar{X}_i^T \hat{\beta} + \hat{v}_{i0}.$$

We use the label (RI) to imply a random intercept model since only the intercept of each small area is random while the other components of β remain fixed.

2.4 Approximation to the Mean Square Error (MSE)

Kackar and Harville (1984) show that, if $\hat{\theta}$ is a translation invariant estimator of θ and the random terms are normally distributed, the mean square error of a predictor of a linear combination of a fixed and random effect can be decomposed into two terms. The first one is due to the variability in estimating the fixed parameters when the components of variance are known, the second term comes from estimating the components of variance.

Since under normality the RIGLS estimator is equivalent to the RMLE estimator and the RMLE is translation-invariant, Kackar and Harville's (1984) results can be applied to the small area means estimators $\hat{\mu}_i$, $i = 1, \ldots, m$:

$$\text{MSE}(\hat{\mu}_i) = E[\hat{\mu}_i - \mu_i]^2 = E[\tilde{\mu}_i - \mu_i]^2 + E[\hat{\mu}_i - \tilde{\mu}_i]^2 \tag{2.5}$$

where $\tilde{\mu}_i$ is the BLUE of μ_i.

The first term of (2.5), that is $\text{MSE}[\tilde{\mu}_i]$, can be obtained by direct calculation as

$$\text{MSE}(\tilde{\mu}_i) = \bar{X}_i^T (G_i^{-1})^T \Omega \bar{X}_i + \sigma^2 \bar{X}_i (G_i^{-1})^T Z_i \left(\sum_{i=1}^{m} Z_i^T G_i^{-1} X_i^T X_i Z_i \right)^{-1} Z_i^T G_i^{-1} \bar{X}_i \tag{2.6}$$

where $G_i = I + \sigma^2 X_i^T X_i \Omega$. Kackar and Harville (1984) point out that the second term of (2.6) is not tractable, except for special cases, and propose an approximation to

it. Prasad and Rao (1990) propose an approximation to this second term and work out the details of their approximation for three particular cases: the random intercept model, random regression coefficient model and the Fay-Herriot model. They also give some regularity conditions for their approximation to be of the second order, and prove that their MSE approximation for the Fay-Herriot model is of the second order. Nevertheless, it seems to be more difficult to give general conditions for more complex models such as model (2.1).

Applying Prasad and Rao's approach, an approximation to the second term of (2.5) is developed in Appendix B.

It is worth noting that the MSE approximation of $(\hat{\mu}_i)$ can be decomposed into three terms:

$$\mathrm{MSE}(\hat{\mu}_i) \approx T_1 + T_2 + T_3 \qquad (2.7)$$

where T_1 and T_2 are respectively the first and the second term of equation (2.6) and T_3 is described in Appendix B.

The term T_1 is the variability of $\hat{\mu}_i$ when all parameters are known, the second term T_2 is due to estimating the fixed effects and the third term T_3 comes from estimating the components of variance.

When sampling fractions are not negligible, estimators of the small area means can be built in the spirit of the finite population approach by predicting specifically for the non-sampled units:

$$\hat{\mu}_i^F = f_i \bar{y}_i + (\bar{X}_i - f_i \bar{x}_i)^T (Z_i \hat{\gamma} + \hat{v}_i) \qquad (2.8)$$

where the superscript F indicates that a correction for the finite population sampling fraction f_i was used; $\bar{x}_i$ is the $(p+1)$ vector of sample means.

The $\mathrm{MSE}(\hat{\mu}_i^F)$ can be obtained by noting that

$$\hat{\mu}_i^F - \bar{Y}_i = (1 - f_i)\left[\left(\bar{X}_i^C\right)^T \left(Z_i(\hat{\gamma} - \gamma) + \hat{v}_i - v_i - \bar{\varepsilon}_i^C\right)\right]$$

where $\bar{X}_i^C = (1 - f_i)^{-1}(\bar{X}_i - f_i \bar{x}_i)$ and $\bar{\varepsilon}_i^C$ is the mean of ε_{ij} for the non-sampled units in the i-th small area. Therefore

$$\mathrm{MSE}(\hat{\mu}_i^F) = (1 - f_i)^2 \left[\mathrm{MSE}^*(\hat{\mu}_i) + N_i^{-1}(1 - f_i)^{-1}\sigma^2\right] \qquad (2.9)$$

where $\mathrm{MSE}^*(\hat{\mu}_i)$ is the equation (2.7) with $\bar{X}_i$ replaced by $\bar{X}_i^C$.

2.5 Estimation of Mean Square Error

It is common practice to estimate the MSE of a linear combination of the fixed and random effects in a mixed model as in (2.1) by replacing estimates of the components of variance respectively in the expression of MSE. This estimator ignores the contribution to MSE due to estimating the components of variance parameters. Several studies (see for example Singh, Stukel and Pfeffermann 1998 or Harville and Jeske 1992) argue that this procedure tends to underestimate the MSE. Prasad and Rao (1990) reported a simulation study which showed that the use of this "naive" estimator leads to severe downwards bias. They also showed for the Fay-Herriot model (a special case of the model (2.1)), using "truncated Henderson" estimates for the variance components, that

$$E(\hat{T}_1) = T_1 - T_3 + o(m^{-1}); E(\hat{T}_2) = T_2 + o(m^{-1});$$

$$E(\hat{T}_3) = T_3 + o(m^{-1}).$$

Harville and Jeske (1992) establish some conditions for the unbiasedness of Prasad and Rao's mean square error estimator. However, considering the more general model (2.1), again it seems more difficult to give general conditions for which the order of bias of Prasad and Rao's estimator is $o(m^{-1})$, especially if iterative procedures as RIGLS are used to obtain the parameter estimates.

Nevertheless, motivated by the simulation study summarised in Section 3.4 and an extensive simulation study described in Moura (1994), we propose to use an estimator similar to Prasad and Rao's for $\mathrm{MSE}(\hat{\mu}_i)$:

$$\hat{\mathrm{MSE}} = \hat{T}_1 + \hat{T}_2 + 2\hat{T}_3. \qquad (2.10)$$

Where $\hat{T}_i$ are obtained from (2.5) by replacing σ^2 and Ω by their respective RIGLS estimators.

From equation (2.9) we can also obtain an estimator for $\mathrm{MSE}(\hat{\mu}_i^F)$ as follows:

$$\hat{\mathrm{MSE}}(\hat{\mu}_i^F) = (1 - f_i)^2 [\hat{\mathrm{MSE}}^*(\hat{\mu}_i) + N_i^{-1}(1 - f_i)^{-1}\hat{\sigma}^2] \qquad (2.11)$$

where $\hat{\mathrm{MSE}}^*(\hat{\mu}_i)$ is the equation (2.10) with $\bar{X}_i$ replaced by $\bar{X}_i^C$.

3. A MODEL-BASED NUMERICAL INVESTIGATION

3.1 Comparison of the Estimators

In order to investigate the properties of alternative estimators, data was used from 38,740 households in the enumeration districts in one county in Brazil. The Head of Household's income was treated as the dependent variable. Two unit level independent variables were identified as the educational attainment of the Head of Household (ordinal scale of 0-5) and the number of rooms in the household (1-11+).

The assumed model is

$$Y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij} \quad i = 1,\ldots,m; j = 1,\ldots,N_i$$

$$\beta_{i0} = \gamma_{00} + v_{i0}; \beta_{i1} = \gamma_{10} + v_{i1}; \beta_{i2} = \gamma_{20} + v_{i2} \qquad (3.1)$$

where x_1 and x_2 respectively represent the number of rooms and the educational attainment of the head of the

household (centred about their respective population means).

The parameter values for the fit model and their respective standard errors are

γ_{00}= 8.456(0.108) γ_{10}= 1.223(0.046) γ_{20}= 2.596(0.086)
σ_{00}= 1.385(0.194) σ_{01} = 0.354(0.66) σ_{02}= 0.492(0.117)
σ_{12} = 0.333(0.054) σ_{11} = 0.234(0.35) σ_{22}= 0.926(0.124)
σ^2 = 47.74(0.345)

To carry out numerical investigations within the model-based framework a simulation was carried out keeping the enumeration district identifiers and the values of the two explanatory variables (X) fixed. Initially the area population means $\bar{X}_{1i}$ and $\bar{X}_{2i}$ were calculated for the whole data set and a randomly selected subsample of 10% of records from each small area was identified. This same subset was retained throughout the simulations (the Simulation subset).

The data generation for the simulations was carried out in two stages using a data generation model which was the General Model (G), the Diagonal Model (D), the Random Intercept Model (RI) as appropriate. In the first case the parameter values were taken from the estimates mentioned earlier. In the second case the off-diagonal terms were set to zero, in the third case only σ_{00} = 1.385 was non-zero.

The first stage of the data simulation process was to generate the level 2 random terms (that is, the non-zero elements of v_{i0} and v_{i1} and v_{i2}) depending on the choice of the data generation model. These random terms were Normally distributed (jointly Normal in the case of the General Data Generation Model and the Diagonal Data Generation Model). At this stage the expected value of the mean for the i-th area conditional on the area level random effects generated by the model m_1 = G, D, RI in the r-th simulation could be obtained:

$$\mu_{im_1}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} \bar{X}_{1i} + \beta_{2i}^{(r)} \bar{X}_{2i}.$$

At the second stage of the data simulation process, unit values (Y_{ij}) were created for each of the data generation models. Having generated the data for the simulation subset under one of the data generation models, all three of the estimation models (G, D and RI) could be fitted to the simulated data to obtain parameter estimates and predictors for the small area means.

For each data generation model m_1 = G, D, RI the whole simulation process was repeated R = 5000 times to yield a set of small area means $\mu_{im_1}^{(r)}$ and predicted means $\hat{\mu}_{im_1,m_2}^{(r)}$, r = 1, ..., R for each small area, i, i = 1, ..., m and for the three estimation models: m_2 = G, D, RI. For each small area and for data generated under model m_1 = G, D, RI, the Mean Square Error (MSE) of the prediction process for each estimation model m_2 may be defined as

$$\mathrm{MSE}[\hat{\mu}_{im_1,m_2}] = R^{-1}\sum_{r=1}^{R}\left(\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)}\right)^2$$

and the absolute relative error (ARE) by

$$\mathrm{ARE}[\hat{\mu}_{im_1,m_2}] = R^{-1}\sum_{r=1}^{R}\left|\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)}\right| / \mu_{im_1}^{(r)}.$$

For comparative purposes we contrast the properties of each estimator with those of the estimator which is the same as the data generation model. Hence we define the Ratio of Mean Square Error (RMSE):

$$\mathrm{RMSE}_{m_2,m_1} = \left\{\sum_{i=1}^{m}\mathrm{MSE}[\hat{\mu}_{im_1,m_2}]\right\} \Big/ \left\{\sum_{i=1}^{m}\mathrm{MSE}[\hat{\mu}_{im_1,m_1}]\right\} \times 100$$

and the Ratio of Absolute Relative Error (RARE):

$$\mathrm{RARE}_{m_2,m_1} = \left\{\sum_{i=1}^{m}\mathrm{ARE}[\hat{\mu}_{im_1,m_2}]\right\} \Big/ \left\{\sum_{i=1}^{m}\mathrm{ARE}[\hat{\mu}_{im_1,m_1}]\right\} \times 100.$$

It will be seen that when the data are generated from a simpler model (*e.g.*, RI) the more complex estimation procedures do not suffer any appreciable worsening of efficiency or bias. On the other hand when the data are generated from a more complex model the simpler estimators have inferior properties. However the difference between the Diagonal and General estimators is much less than between these and the Random Intercept Estimator. From Table 1 one would conclude that it is worth introducing additional random coefficients of some kind, beyond the simple Random Intercept model assumptions, but not necessarily the full General Model.

Table 1
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the three Estimators and Three Data Generation Models

Estimator	Data Generation Model		
	G	D	RI
General (G)	100.0 (100.0)	101.8 (100.9)	101.2 (100.6)
Diagonal (D)	108.8 (82.6)	100.0 (100.0)	100.2 (100.1)
R. Intercept (RI)	131.9 (176.9)	109.1 (105.6)	100.0 (100.0)

The summary measures in Table 1 are average properties over all small areas. A careful analysis of the MSE performance of the estimators for each small area shows that there is a modest increase in the MSE for the Diagonal Estimator

compared to the General Estimator for all areas, whereas for the Random Intercept estimator a relatively small number of areas exhibit a substantial increase in MSE. A similar pattern occurs between the Diagonal and Random Intercept estimator when the Diagonal Data Generation Model is used.

3.2 Introducing a Small-Area Level Covariate

In this section an attempt is made to investigate the impact on small area estimates of introducing an area covariate Z. Unfortunately for the data set used, it was not possible to identify a single contextual area level covariate which had a substantial effect on the multilevel models. Nevertheless, the number of cars per household in each small area was a useful covariate for the random coefficients for the individual level random slopes coefficients for "Room" and "Edu", but not for the random intercept term. This was observed after some preliminary model fit analysis on the real data. Although the "numbers of cars" was the best small area level covariate found to explain between area variation, it was not as powerful at the individual level as "Room" and "Edu", the individual level covariates chosen.

The model above with the small area covariate Z can be written as

$$Y_{ij} = \beta_{i0}+\beta_{i1}x_{1ij}+\beta_{i2}x_{2ij}+\varepsilon_{ij} \quad i = 1,\ldots, m; j = 1,\ldots,N_i$$

$$\beta_{i0} = \gamma_{00}+\nu_{i0};\ \beta_{i1} = \gamma_{10}+\gamma_{11}z_i+\nu_{i1};\ \beta_{i2} = \gamma_{20}+\gamma_{21}z_i+\nu_{i2}. \quad (3.2)$$

The small area random effects were assumed uncorrelated in order to avoid convergence failure in the simulation study.

Table 2 reports the parameter estimates and their respective standard errors obtained by fitting the Diagonal Model with the Z covariate (3.2) and without the Z covariate (2.1). It is worth noting the significant reduction of all the components of variance estimates, except $\hat{\sigma}_{00}$ and $\hat{\sigma}^2$, after introducing the explanatory area covariate Z.

In order to investigate the effect of misspecification of the Z variable, the model based simulation procedure described in section 3.1 was applied to the two models above, where the data generation was done according to the parameters presented in Table 2. Table 3 summarises the simulation results.

It is worth noting that in both cases there is a significant loss of efficiency by using an unsuitable estimator. It can also be seen from an individual analysis of MSE for each small area that a considerable gain in efficiency is achieved with the introduction of a small area covariate Z over the diagonal model. For many small areas the MSE of the Diagonal with Z is significantly less than the MSE of the corresponding estimator without Z. Even for those few areas in which the MSE of the Diagonal with Z is unchanged or even slightly increased by the introduction of Z, the difference is not appreciable.

Table 2
Parameter Estimates and Standard Errors for General Model with Area Level Covariate: Demographic Data

Parameter	Diagonal Model with Z	Diagonal Model
γ_{00}	8.442(0.112)	8.688(0.136)
γ_{10}	0.451(0.179)	1.321(0.085)
γ_{20}	0.744(0.272)	2.636(0.134)
γ_{11}	3.779(0.507)	-
γ_{22}	1.659(0.323)	-
σ_{00}	0.745(0.308)	0.637(0.303)
σ_{11}	0.237(0.083)	0.471(0.116)
σ_{22}	0.700(0.197)	1.472(0.295)
σ^2	44.00(1.05)	44.01(1.05)

Table 3
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the Diagonal and the Diagonal with Z Estimators Under the Two Respective Data Generation Models

Estimator	Data Generation Model	
	Diagonal	Diagonal with Z
Diagonal	100.0 (100.0)	110.3 (125.4)
Diagonal with Z	126.2 (107.5)	100.0 (100.0)

3.3 Comparisons with Regression Estimator

One essential advantage of the multilevel models over regression models is to recognize that groups (here the small areas) share common features; they are not completely independent as could be assumed, for example by using separate linear regression model for each small area. Nevertheless, the relatively small intraclass correlation observed for the data set used plus the fact that each small area has on average 28 units, could make one think that in this case the use of the multilevel model would not result in great improvement in the small area estimators. However, it is gratifying to know that even in these circumstances the multilevel model small area estimator performs on average better than the synthetic separate regression estimator, under either the multilevel model or even under the regression model. Table 4 illustrates this finding.

The multilevel data generation model used was the General one with the parameters given in section 3.1. The parameters used in the data generation regression model were obtained by fitting a separate regression for each small area.

It can be seen from Table 4 that the Separate Regression estimator which does not explore the difference of small areas through small area random effects shows substantial loss of efficiency when compared with the General estimator.

Table 4
Ratios of Mean Square Errors and Ratios of Absolute Relative Errors (in parentheses) for the General and the Separate Regression Estimators Under the Two Respective Data Generation Models

Estimator	Data Generation Model	
	General	Separate Regression
General	100.0 (100.0)	88.1 (83.1)
Separated Regression	247.6 (154.7)	100.0 (100.0)

Figure 1 illustrates this fact by showing a plot of the ratio of mean square error between the General estimator and the Separate Regression estimator for each small area. To demonstrate the effect of the small area sample size on the efficiency, the ratio of the MSEs is plotted against the sample size for each small area. It is clear from Figure 1 that the gain in efficiency tends to decrease as the sample size increases.

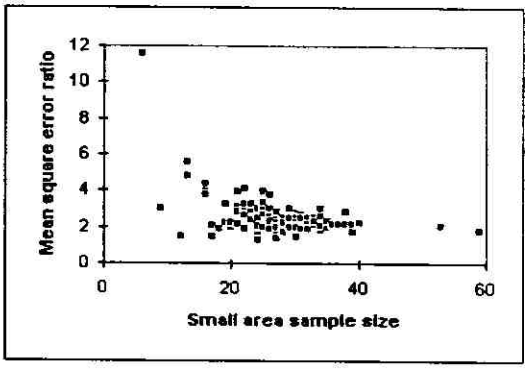

Figure 1. Model-based efficiencies of the general estimator compared with the separate regression estimator for each small area

3.4 An Evaluation of the MSE Approximation and the MSE Estimator

From the simulation results we may investigate the properties of the MSE approximation (2.7). If we consider the General estimator when the General Data Generation model is used the MSE approximation appears to be very good. The average underestimation of the MSE approximation was 0.31% of the MSE value with a range from the largest underestimate of 5.4% of the MSE value through to a largest overestimate of 4.8% of the MSE value. For the situation considered here T_1 contributed on average 94.6% of the total variation and T_3 a further 4.3%. Given the large component of variance due to σ^2, these results are not unexpected. For individual areas the component T_1 varied between 87.4% to 99.1% of the total and T_3 varied between 0.7% and 10.5% of the total. The component T_2 never contributed more than 2.2% of the total MSE for any area.

We also investigated the performance of the MSE estimator represented by equation (2.10) against the "naive" estimator of the MSE, which does not consider the last term of (2.10). The average Root Mean Squared Error of the proposed MSE estimator is 17.5% ranging from 4.7% to 32.3%, while for the naive estimator the average is 20.9% ranging from 5.2% to 47.5%. The MSE estimator is on average unbiased while the naive MSE estimator underestimates the MSE on average by 9.1%, its relative bias ranging from -23.5% to -0.9%. Our results agree with others, see Singh, Stukel and Pfeffermann (1998) and Prasad and Rao (1990), which show that the naive estimator can exhibit severe bias.

4. DISCUSSION

Prasad and Rao (1990) and Battese *et al.*, (1981, 1988) have demonstrated that models which include small area specific components of variance can provide greatly improved small area estimators. Some of the numerical results in this paper show that within the model-based simulation framework even better estimators can be obtained by allowing the small area slopes as well as the intercept to be random.

The overall conclusions from this investigation for this set of parameter values are that: a component of variance model more complex than the Random Intercept estimator is beneficial; overspecification of the model (*e.g.*, using the General estimator with data generated under the Random Intercept Model) does not lead to serious loss of efficiency; the use of small area covariates can also improve the small area estimates; and the use of multilevel models should be preferred rather than the Separate Regression Model. The simulation study confirms that the MSE approximation appears to be precise and the MSE estimation is approximately unbiased, reflecting the variation in MSE between areas, but further theoretical investigation about the exact order of the approximation should be done.

Clearly model fitting and diagnostics are crucial. If we apply a general mixed model in circumstances where it is only a poor fit to the data, then the results may be disappointing. Considerably more investigation is needed to understand what characteristics of specific small areas are likely to provide efficiency gains if general mixed models are used rather than simpler models.

ACKNOWLEDGEMENTS

We would like to thank the referees and the Editor for their helpful comments on the earlier version of this paper.

APPENDIX A: RESTRICTED ITERATIVE GENERALIZED PROCEDURE

The generalized least squares estimator of γ in the model (2.1) is given by

$$\tilde{\gamma} = (Z^T X^T V^{-1} XZ)^{-1} (Z^T X^T V^{-1} Y) = \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} X_i Z_i \right)^{-1} \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} Y_i \right) \quad (A.1)$$

where $V = \text{Diag}(V_1, ..., V_m)$ and $V_i = \sigma^2 I + X^T \Omega X$ is the covariance matrix of Y_i, $i = 1, ..., m$.

However, V is assumed to be a function of unknown parameters, thus γ cannot be estimated using (A.1). On the other hand, if γ is known then

$$Y^* = \text{vech}\,[(Y - XZ\gamma)(Y - XZ\gamma)^T] \quad (A.2)$$

is an unbiased estimator of vech(V). Furthermore vech(V) is a linear function of θ. Then we can consider the following linear model:

$$Y^* = F\theta + \xi. \quad (A.3)$$

Where $F = \partial \text{vech}(V)/\partial\theta$ and ξ is a random variable with mean $O = (0, ..., 0)$ and the covariance of ξ is given by $V_\xi = 2\varphi_n(V \otimes V)\varphi_n^T$. The matrix φ_n is any linear transformation of vec(A) into vech(A), and A is any $n \times n$ matrix such that vech(A) = $\varphi_n$vec(A), see Fuller(1987) for further details. Then, assuming that F has full rank and V_ξ is known and non-singular, it may be shown that the Generalized Least Square Estimator of θ is given by

$$\tilde{\theta}_a = \text{cov}\,(\tilde{\theta}_a) \left(\frac{\partial \text{vec}(V)}{\partial\theta} \right)^T \left(\frac{1}{2} V^{-1} \otimes V^{-1} \right) \text{vec}(\tilde{Y}\tilde{Y}^T) \quad (A.4)$$

where

$$\text{cov}\,(\tilde{\theta}_a) = \left[\left(\frac{\partial \text{vec}(V)}{\partial\theta} \right)^T \left(\frac{1}{2} V^{-1} \otimes V^{-1} \right) \left(\frac{\partial \text{vec}(V)}{\partial\theta} \right) \right]^{-1}$$

and

$$\tilde{Y} = Y - XZ\gamma.$$

Note that $\tilde{\theta}_a$ depends on θ and γ, so both may be iteratively estimated. The IGLS procedure starts with an initial estimate of V (that is, setting initial values of θ) which produces an estimate of γ. Hence replacing the initial estimate of V together with the estimate of γ in (A.1) provides an improved estimate of θ. In most cases convergence is achieved after a few iterations between equations (A.1) and (A.4), although it is not always guaranteed.

The RIGLS approach is based on the fact that if γ is estimated by using generalised least squares with V known then

$$E\left[(Y - XZ\hat{\gamma})(Y - XZ\hat{\gamma})^T\right] = V - XZ(Z^T X^T V^{-1} XZ)^{-1} Z^T X^T.$$

The equation above suggests that we use

$$(Y - XZ\gamma)(Y - XZ\gamma)^T + XZ(Z^T X^T V^{-1} XZ)^{-1} Z^T X^T \quad (A.5)$$

instead of $(Y - XZ\gamma)(Y - XZ\gamma)^T$ at each iteration cycle described above in order to obtain an approximately unbiased estimator of V and consequently of θ.

As pointed out by Goldstein (1986, 1989), if we start with a consistent estimate of γ, say the ordinary least squares estimator, then the final estimates will be consistent providing finite fourth moments exist.

It is worth noting that it is possible for the above procedure to yield negative estimates of variances. This problem can be avoided by imposing constraints at each iteration. For further details on this issue see Goldstein (1986).

APPENDIX B: AN APPROXIMATION TO $E[\hat{\mu}_i - \tilde{\mu}_i]^2$

Prasad and Rao (1990), based on Kachar and Harville (1984), developed a second order approximation to the second term of (2.5) under some regularity conditions:

$$E[\hat{\mu}_i - \tilde{\mu}_i]^2 \approx T_3 = tr\left[\left(\frac{\partial d_i}{\partial\theta} \right) V \left(\frac{\partial d_i}{\partial\theta} \right)^T E\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \right] \right] \quad (B.1)$$

where, for the model (2.1), $d_i = \bar{X}_i^T K_i (I \otimes \Omega) X^T V^{-1}$, $K_i = [0, ..., I, ... 0]$, is the $(p+1) \times (p+1)m$ matrix with the identity matrix I of order $p+1$ in the i-th position and 0 as the null matrix of order $p+1$, and $\hat{\theta}$ is any translation-invariant estimator of $\theta = (\theta_1, ..., \theta_s)$ where $\theta_s = \sigma^2$ and $\theta_k; k = 1, ..., s-1$ are the distinct elements of Ω. Goldstein (1989) proves that under normality of the random terms of model (2.1), the RIGLS estimator of θ is equivalent to the Restricted Maximum Likelihood Estimator (RMLE), which is translation invariant.

Let us approximate $E\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right]$ to the asymptotic covariance matrix of the RMLE estimator (B). The jk-th element of B^{-1} is given by(see Harville (1977))

$$b_{jk}^* = Tr\left(\sum_{i=1}^{m} P_i \frac{\partial V}{\partial\theta_j} P_i \frac{\partial V}{\partial\theta_K} \right)$$

for j and $k = 1, ..., s$ where $P_i = V_i^{-1} - V_i^{-1} X_i Z_i (\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} X_i Z_i) \; Z_i^T X_i^T V_i^{-1}$. Let $b_{j,k}$ be jk-th element of B. After some matrix algebra, it can be shown

that

$$T_3 = \bar{X}_i^T (G_i^{-1})^T \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \Delta_j C_i \Delta_k^T \right) G_i^{-1} \bar{X}_i -$$
$$2\bar{X}_i^T (G_i^{-1})^T \left(\sum_{j=1}^{s-1} b_{j,s} \Delta_j \right) R_i \Omega \bar{X}_i + b_{ss} \bar{X}_i^T \Omega S_i \Omega \bar{X}_i \quad \text{(B.2)}$$

where $C_i = \sigma^{-2} G_i^{-1} X_i^T X_i$; $R_i = \sigma^{-4} G_i^{-2} X_i^T X_i$; $S_i = \sigma^{-6} G_i^{-3} X_i^T X_i$; and

$$\Delta_k = \frac{\partial \Omega}{\partial \Omega_k} \; k = 1, \ldots, s-1$$

is the $s-1$ square derivative matrix with respect to θ_k; $k = 1, \ldots, s-1$.

REFERENCES

BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

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.

FULLER, W.A. (1987). *Measurement Error Models*. Chichester: John Wiley.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares estimation. *Biometrika*, 73, 43-56.

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76, 622-623.

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

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

HARVILLE, D.A. (1977). Maximum likelihood approach to variance component estimation and related problems. *Journal of the American Statistical Association*, 72, 320-340.

HARVILLE, D.A., and JESKE, D.R. (1992). Mean squared error of estimation on prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.

HENDERSON, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.

HOLT, D., and MOURA, F. (1993a). Mixed models for making small area estimates. In: *Small Area Statistics and Survey Design*, (G. Kalton, J. Kordos, and R. Platek, Eds.) 1, 221-231. Warsaw: Central Statistical Office.

HOLT, D., and MOURA, F. (1993b). Small area estimation using multilevel models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-31.

KACKAR, R.N., and HARVILLE, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

LONGFORD, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika*, 79, 817-827.

MOURA, F.A.S. (1994). Small Area Estimation Using Multilevel Models. University of Southampton. Unpublished Ph. D. Thesis

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

SINGH, A.C., STUKEL, D., and PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B. 377-396.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 91-98
Statistics Canada

A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France

GEORGES DECAUDIN and JEAN-CLAUDE LABAT[1]

ABSTRACT

Since France has no population registers, population censuses are the basis for its socio-demographic information system. However, between two censuses, some data must be updated, in particular at a high level of geographic detail, especially since censuses are tending, for various reasons, to be less frequent. In 1993, the Institut National de la Statistique et des Études Économiques (INSEE) set up a team whose objective was to propose a system to substantially improve the existing mechanism for making small area population estimates. Its task was twofold: to prepare an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is reported on here, is flexible and reliable, without being overly complex.

KEY WORDS: Population estimates; Administrative files; Robust estimation.

1. INTRODUCTION

In France, as in all countries that do not have population registers, censuses of the population are the cornerstone of the socio-demographic information system. However, censuses are quite massive operations that cannot at present be carried out more often than once every seven or eight years. In the interval between censuses, it is therefore necessary to update some information, especially at a high level of geographic detail, particularly since for various reasons, censuses are tending to be less frequent. Thus, small area population estimates are a major challenge for the Institut National de la Statistique et des Études Économiques (INSEE).

Despite the progress achieved in this field, the situation in 1993 still seemed fairly unsatisfactory. When figures from the 1990 population census were compared to the population estimates made on the basis of the previous census (1982) for the metropolitan departments, the differences noted were sometimes sizable.

INSEE therefore created a methodology team whose mission was to propose a system that would substantially improve the existing mechanism. Initially, the next census was to take place in 1997. It therefore seemed reasonable to have the new system operate on an experimental basis until the census, so as to see how well it worked before using it in actual production. When the census was postponed to 1999, it became more necessary to bring the project to a successful conclusion quickly, so as to be able to use the new system in 1996.

To achieve its objective, the team devoted itself, with maximum pragmatism, to a twofold task: to develop an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is described here, is not overly complex and seems effective. A more detailed description of it is provided in Decaudin and Labat (1996).

2. MAIN CONCLUSIONS

The team's main conclusions are as follows:

1) It is impossible to improve total population estimates using sample surveys, unless the survey is conducted on such a scale that it would be similar to a census.
2) No single administrative source adequately reflects changes in the population. At the local level, all sources can exhibit drift, breaks, jolts, *etc.*, which are not always easy to detect. Furthermore, even at the local level, it is often quite difficult if not impossible to get the agency responsible to provide explanatory details, much less corrections in the case of errors. In any event, it is unwise to rely on a single administrative source, however good it may be, since its permanency is never guaranteed.
3) On the other hand, total population estimates can be improved substantially by simultaneously using several sources. A "multi-source" system, similar to the one presented here but more rudimentary, was tested retrospectively over the intercensal period 1982-1990, for the 96 metropolitan departments. The mean error (mean deviation as an absolute value from the results of the March 1990 census) fell below 0.9%, whereas the mean error registered at the time, with the estimation system then in place, was 1.4%.

3. SIMULTANEOUS USE OF SEVERAL SOURCES

For using several sources jointly, different methods are possible.

A method that is universal – and easy to implement – is multiple regression. In simplified form, this amounts to using, for any area z, the following relationship:

$$P(n+1,z)/P(n,z) = c + \sum_S (k_S N_S(n+1,z)/N_S(n,z)),$$

[1] Georges Decaudin and Jean-Claude Labat, Institut National de la Statistique et des Études Économique, 18, Blvd. Adolphe-Pinard, 75765 Paris, CEDEX 14.

where $P(n, z)$ is the population of area z on January 1 of year n, the values $N_S(n, z)$ are the numbers from each source S on the same date and k_S are coefficients, which are estimated by multiple regression over a past period. Here c is a constant term that is used only in the regression, with calibration on the national population serving to correct any drift.

This method is used in various countries, including Canada and the United States (for example, see Statistics Canada 1987 and Long 1993). Nevertheless, it was not adopted because it has numerous drawbacks:

- it must be possible to estimate the coefficients, which requires data from each source extending back over a fairly long period;
- the coefficients can change over time, without it being possible to control this change;
- as noted above, the administrative sources are, for various reasons (changes in regulations, abrupt shifts in management, errors, *etc.*), subject to what might be called "anomalies". For each source S, the scope of these anomalies is reflected in part in the coefficient k_S, to an extent that depends on how great their medium-term effect has been over the calibration period [la période d'étalonnage]; but anomalies nevertheless occur in estimates with the same weight as the "good" data from the same source. The estimates are then highly distorted.

Another method is known as the "*composite*" method. Each source is used to estimate the population in one or more age classes: age class X, which is well-covered by the source, but also sometimes another class that definitely exhibits a pattern very similar to that of class X (for example, the "30-45" age group, if X represents the "under 18" age group). It is then necessary to have appropriate indicators for the other components of the population and correctly manage the consolidation of these estimates "in parts".

This type of method, used in the United States (Long 1993), seemed to us to be problematic, especially because of the difficulty of adequately dealing with "anomalies".

The *proposed "multi-source" system* is based on a robust synthesis of estimates from different sources. It combines demographic reasoning with purely statistical techniques. It draws on the experiments conducted by the INSEE's regional directorate in Brittany in the early 1970s (Laurent and Guéguen 1971; Guéguen 1972). Should one of the sources fail, such a system is not prevented from functioning, even though its performance may be somewhat diminished.

4. DEMOGRAPHIC BASE

The demographic reasoning which is at the base of the system is elementary: assuming that we know the total population $P(n)$ for an area on January 1 of year n, the population $P(n + 1)$ of the area on January 1 of year $n + 1$ is deduced by summing the two components of the change during year n: natural increase (births minus deaths), and net migration (immigrants minus emigrants).

$$P(n + 1) = P(n) + N(n) - D(n) + I(n) - E(n).$$

In France, natural increase data are provided annually at the commune level by vital statistics. If the latter are not yet available in final form, which is often the case in the third quarter of year $n + 1$, it is easy to estimate them with a low margin of uncertainty.

The only unknown, then, is net migration for year n: $SM(n) = I(n) - E(n)$ or what amounts to the same thing, the net migration rate $T(n) = SM(n)/P(n)$. In other words, estimating the population comes down to estimating net migration since the last date on which the population is known (or is assumed to be known), and vice versa.

In France, net migration figures are of some importance, although less so than in other countries such as Canada or the United States. In addition, they generally exhibit a certain inertia, at least at relatively aggregated geographic levels. One way to assess the influence of changes to them from one intercensal period to the next is to measure the errors that would have been committed during each period if the population had been estimated by using the average annual net migration rates for the preceding period. Over the period 1982-1990, for the departments (excluding Corsica), the mean end-of-period error (in 1990, at the end of eight years) would have been only 1.3%. It was not certain, when the team started its work, that much greater accuracy could be achieved. However, both in 1975 and in 1982, the mean error that would have been committed with the trend method would have been much greater: 2.8% and 2.7% respectively (over seven years). It would therefore seem that the period 1982-1990 was exceptional and that in the future the difference will again be more pronounced.

5. ESTIMATES FROM THE DIFFERENT SOURCES

From each source, using an appropriate method, we draw an estimate of annual net migration rate for the population as a whole. The methods that may be used depend on the data available.

For each of the sources tested and found to be "good", at least at the departmental level, a method is proposed. The five sources retained are the following: housing tax; electrical utility customers; children receiving family allowances; educational statistics; electoral file.

The data on the composition of households for tax purposes, which appear in the income tax files, are the sixth source that should provide very good results. However, to date, these data have been analysed for only a few departments, and the methodology for using them is not yet completely defined.

We also propose to integrate a trend estimate of the net migration rate into the system.

Two categories of methods are used. The first concerns the sources relating to households; the second concerns those relating to individuals.

5.1 Sources Relating to Households

Some sources provide information on changes in the number of households. This is the case with the files on *housing taxes* (HT) and *electrical utility customers* (EUC). The housing tax is one of the four main local direct taxes. As its name indicates, it applies to occupied dwellings, with main residences and secondary residences being treated separately. The housing tax file takes account of the situation on January 1 of the taxation year. Starting in the 1980s, the HT source was the basis for the departmental population estimates developed by INSEE (Descours 1992). In the early 1990s, it was replaced by the EUC source, in light of the distortions caused by a change to the HT management system which gradually worked its way through all departments.

The method adopted for using these sources follows classical principles. It leads directly to an estimate of the total population, and it involves three main stages:

1) estimating the number of households;
2) estimating average household size and from there, estimating the population of households;
3) adding the "non-household" population.

In the first stage, it is assumed that the number of households changes in accordance with the data supplied by the source (number of main residences for HT purposes or number of electrical utility customers). The second stage is more delicate. It is based on both the use of statistics on dependants from the HT files and on a trend estimate of average household size.

In the proposed "multi-source" system, we move on to the net migration rate, for comparison with other sources, using vital statistics data (*cf.* Section 4).

5.2 Sources Relating to Individuals

The other sources used concern individuals. Only a certain age group X of the population is generally covered adequately. The method then involves two main stages:

1) estimating, from the source, the net migration rate for the population aged X;
2) from there, estimating the net migration rate for the population as a whole.

The second stage is based on the following statistical relationship, observed in the past, between the change, from one period to another, of the overall net migration rate (T) and the change in the net migration rate for the population aged X (TX):

$$T_2 - T_1 = \delta_X (TX_2 - TX_1),$$

where δ_X is a coefficient close to 1, depending on the age group X. This relationship is similar to the one used by de Guibert-Lantoine (1987) to estimate the population on the basis of educational statistics.

For the corresponding age groups in the different sources used, the values, estimated by linear regression, of the coefficient δ_X (+/-2 standard deviations) are shown in tables 1 and 2.

Table 1
Estimates of δ_X on Departments, Excluding Corsica, Internal Net Migration

Period 1	Period 2	Age at end of period		
		0-19	10-14	35 and over
1962-1968	1968-1975	0.76 (+/- 0.04)	0.69 (+/-0.06)	1.24 (+/-0.09)
1968-1975	1975-1982	0.77 (+/-0.03)	0.88 (+/-0.06)	1.56 (+/-0.08)
1975-1982	1982-1990	0.70 (+/-0.11)	0.49 (+/-0.10)	1.26 (+/-0.17)

Table 2
Estimates of δ_X Over the Two Periods 1975-1982 and 1982-1990, Excluding Corsica, Total Net Migration

	Age at end of period		
	0-18	9-15	35 and over
Departments	0.65 (+/-0.11)	0.57 (+/-0.10)	1.22 (+/-0.16)
Department – employment zone	0.65 (+/-0.04)	0.59 (+/-0.04)	1.17 (+/-0.06)

The approach followed in the first stage depends on the source:

Electoral File

Annual migration figures for voters in the selected age group (30 and over) are supplied directly by the electoral file managed by INSEE. We go from the rate of net migration of voters to the residential net migration rate by dividing the former by a coefficient reflecting the magnitude of the change in the electoral file.

Educational Statistics

The net migration figure for those in the 5-9 age group is obtained by subtracting their number in year n from that of the same cohorts the next year (that is, from those in the 6-10 age group in year $n + 1$) and deducting deaths.

Children Receiving Family Allowances

The number of persons in the 0-17 age group is estimated on the assumption that it evolves similarly to the number of children receiving family allowances. From this a figure for the net migration of young persons is obtained by comparing this estimate to a hypothetical change in the youth population without migration, that is, a change due solely to natural increase.

6. SYNTHESIS

6.1 Principles

The different basic estimates of the annual net migration rate are treated statistically in order to obtain a "synthetic rate", to be used as the final estimate. The treatment serves to eliminate outliers, underweight suspect values and, more generally, assign to each source a weight that reflects its performance.

More specifically, since each source can "drift", the different basic estimates are generally biased; they are first corrected for the national bias of the corresponding source for the year considered, a bias that is estimated in advance. In proceeding in this way, we implicitly assume that the difference between the local bias and the national bias is minor in relation to the irreducible unexplained portion of the difference (flou irréductible). Once we have estimates for a number of years, it should be possible to test this hypothesis and if necessary, replace it with one that corresponds more closely to reality, so as to improve the correction of biases at the local level.

It should be noted that such a seemingly simple operation as correcting the national bias nevertheless requires several precautions. The solution that consists in carrying out a gross calibration on the national net migration rate, considered by definition as a good reference, is not very satisfactory, owing to anomalies that may distort the calibration. It is therefore preferable to estimate the biases by means of a process in which we also eliminate anomalies. The process is similar to the one used for synthesis, which is described below. However, the determination of biases, assumed to be national in scope and therefore calculated for 96 departments, is less sensitive to anomalies than the determination of synthetic rates, calculated over a small number of sources. Only major anomalies are likely to significantly throw off the calibration of the rates and must therefore be corrected.

The "synthetic" net migration rate is a weighted mean of the basic estimates thus calibrated. Each source S is assigned an initial weight W_S that is supposed to reflect its medium-term accuracy. But in addition, for a given year and area, this weight is modulated to take account of the plausibility of the corresponding rate. Thus, if a rate is "abnormally distant" from the rates obtained from other sources – in practice, from a central value for all rates for the area – its weight is cancelled or reduced. For this, we look at the distance between the rate obtained from each source and the central value identified, and we compare it to a "norm" of distance NO_S specific to the source, determined empirically on the basis of the data available: if the distance is less than "a times the norm", the weight is not automatically changed; if it is greater than "b times the norm", it is set at 0; between the two, the weight is multiplied by a coefficient, included between 0 and 1, calculated by interpolation.

Note that the trend estimate is formally treated like those from exogenous sources; its weight is cancelled when it is considered as implausible because it is too far from the other estimates.

The synthesis is achieved automatically, which ensures homogeneity and an explicit logic to the treatments carried out. This does not, however, eliminate the need to control the results obtained.

6.2 Theoretical Presentation

On the theoretical level, we sought to use reasonings and robust estimation techniques, such as described in Hoaglin, Mosteller and Tukey (1983). The method adopted falls within the framework of M-estimators of central tendency and more specifically in the category of W-estimators, which use the reweighted least squares algorithm.

Since the net migration rates for year n and area z obtained from different sources S (and corrected for their national biases) are denoted $TC_S(n, z)$, the synthetic rate $T(n, z)$ solves the implicit equation:

$$\sum_S W_S \cdot NO_S \cdot \Psi\left(\frac{TC_s(n,z) - T(n,z)}{NO_S}\right) = 0,$$

where the function Ψ is of the type that redescends to a finite rejection point:

$$\Psi(r) = r \qquad \text{for } |r| \le a,$$

$$\Psi(r) = r\frac{b - |r|}{b - a} \qquad \text{for } a < |r| \le b,$$

$$\Psi(r) = 0 \qquad \text{otherwise.}$$

Using an iterative process, we can gradually refine the automatic processing of suspect data.

6.3 First Analysis of the Distances From Each Rate to the Central Value for the Rates

1) For each area z we calculate a first central value of the "calibrated" rates $TC_S(n, z)$. The central value used must not be overly sensitive to the possible existence of quite distant values for some sources, but at the same time it must be influenced by a source to the extent that the source is on average more accurate. Under these conditions, rather than choosing the median – which would meet the first condition – we use a statistic of rank that is a little more elaborate but nevertheless simple, owing to the small number of values; this statistic is the mean, weighted by respectively 1/2, 1/4, 1/4, of the three quartiles:
   - the median of the rates $TC_S(n, z)$ weighted by the initial weights W_S,
   - the lower quartile (Q1) of the weighted rates,
   - the upper quartile (Q3) of the weighted rates.

2) The rates $T1(n, z)$ thus obtained are calibrated on the net migration rate for the higher level, by simple translation:

$$TC1(n, z) = T1(n, z) + TREF(n) - \sum_z (T1(n, z)P(n, z)) \Big/ \sum_z P(n, z)$$

where $P(n, z)$ is the population of area z on January 1 of year n and $TREF(n)$ is the net migration rate for the higher level (the national rate for the departmental synthesis).

3) For each area, we calculate the differences between each rate and this calibrated central value:

$$EC1_S(n, z) = | TC_S(n, z) - TC1(n, z) |.$$

4) For each source and each area, the size of this difference is assessed in relation to the "norm" of distance NO_S specific to the source. This "norm" is determined empirically on the basis of the available data: theoretically it is the average of the distances observed in the past, excluding anomalies. The result is a first modulation of the weight originally assigned to this source:
   - if $EC1_S(n, z) \leq a1\, NO_S$, where $a1$ is a parameter to be chosen (in the vicinity of 2), we do not change W_S, the initial weight for S. In other words, if $WM1_S(n, z)$ is the modulation coefficient of W_S (coefficient included between 0 and 1), we take $WM1_S(n, z) = 1$;
   - if $EC1_S(n, z) > b1\, NO_S$, where $b1$ is another parameter (in the vicinity of 3), we set W_S at 0, meaning that we eliminate source S: $WM1_S(n, z) = 0$;
   - if $a1\, NO_S < EC1_S(n, z) \leq b1\, NO_S$, we interpolate $WM1_S(n, z)$ as a function of the value of $EC1_S(n, z)$:

$$WM1_S(n, z) = (b1\, NO_S - EC1_S(n, z))/((b1 - a1)NO_S).$$

5) At the end of this first phase, we therefore have new weights specific to each source and each area, which would allow us to locally eliminate or underweight suspect rates: $W1_S(n, z) = W_S WM1_S(n, z)$.

6.4 Iterations

1) Using the weights thus modified $W1_S(n, z)$, we estimate a new central value for each area, this time taking the weighted average of the rates:

$$T2(n, z) = \sum_S (TC_S(n, z) W1_S(n, z)) \Big/ \sum_S W1_S(n, z).$$

2) We calibrate each rate $T2(n, z)$ on the net migration rate for the higher level, by translation. We obtain $TC2(n, z)$.

3) We calculate, in each area, the differences between each rate and the calibrated average rate: $EC2_S(n, z) = | TC_S(n, z) - TC2(n, z) |$. Using these differences, we calculate new modulation coefficients for the initial weights, using the parameters $a2$ and $b2$, which may be different from $a1$ and $b1$ (theoretically they would be lower). We thus obtain new weights $W2_S(n, z)$ which more effectively take account of anomalies, since the latter are assessed in relation to a better central tendency. With these weights, we estimate a new synthetic rate $T3(n, z)$, which is calibrated on the higher level to obtain $TC3(n, z)$.

4) The operations described in point 3 are repeated with the same parameters $a2$ and $b2$. The tests conducted at the departmental level over the period 1982-1990 show that the convergence is generally rapid; the rates are quite often stabilized by the fourth iteration.

7. IMPLEMENTATION AT THE DEPARTMENTAL LEVEL

The estimation system outlined above, which is operationalized for 1990 and subsequent years, was implemented by the project team for the year 1990 at the departmental level, with the following five sources: housing tax (HT), electrical utility customers (EUC), family allowances (FA), educational statistics (ES), electoral file (EF), plus the trend estimate (TREND).

Figure 1 shows the results obtained for several departments. Table 3 shows the values of the weights and norms used to make the system operate. This table also shows certain statistics obtained from the synthesis of the net migration rates; in particular they concern the differences between the rates obtained from each source and the synthetic rates.

Table 3
Implementation for Year 1990 at Department Level
Parameters and Statistics

	HT	EUC	FA	ES	EF	TEND
Weight	115	100	80	70	80	100
Norm	0.15	0.17	0.19	0.20	0.19	0.12
Number of rates	96	96	89	96	94	96
Average distance	0.55	0.14	0.30	0.19	0.14	0.13
Number of "aberrant" rates	37	2	17	3	1	6
Average of distances without "aberrant" rates	0.15	0.13	0.16	0.16	0.13	0.11

Note: - Coefficients (a; b) applied to norms: (2,5; 3,5) in the first iteration, then (2; 3).
- The values of the distances and norms correspond to rates expressed as a %.
- Distances are calculated in relation to the synthetic rates after three iterations.
- "Aberrant" rates are those for which the weight is cancelled after three iterations.

The results suggest that the system is even more effective than indicated by the summary retrospective test carried out on the 1982-1990 intercensal period with the same sources. Aside from the HT source, which is still distorted, the estimates from the different sources are more convergent than they were on average in the retrospective test (see Table 4).

There is nothing surprising about this, given the rudimentary state of the system tested on the 1982-1990 intercensal period. The data used were rough or even

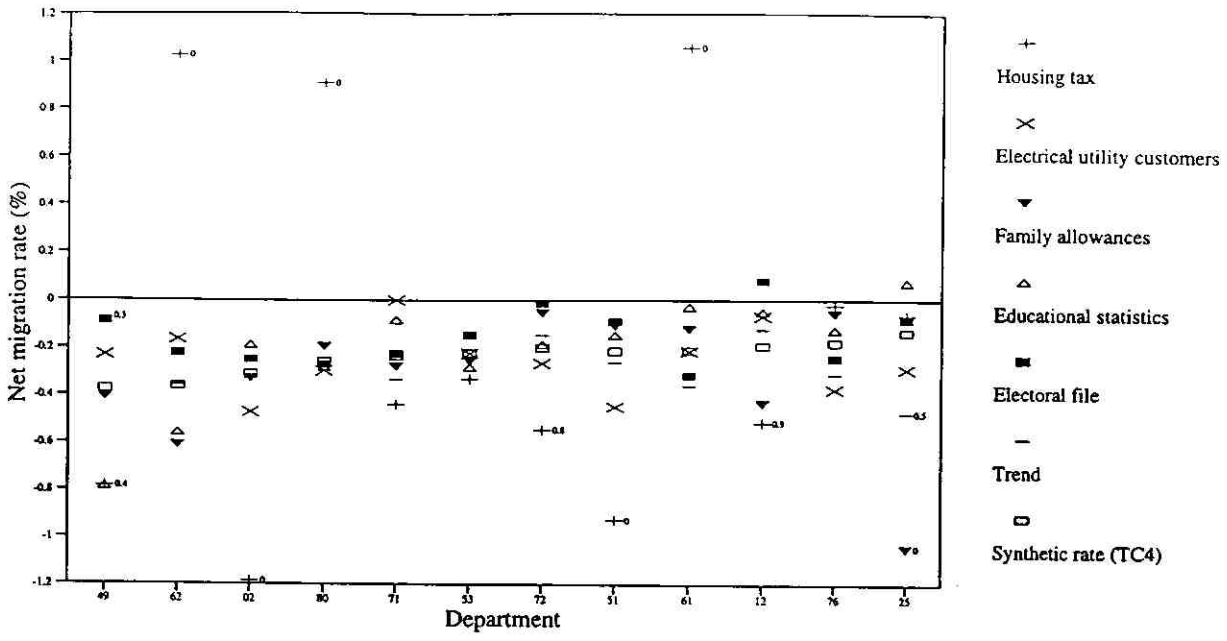

Figure 1: Summary of Net Migration Rates for 1990 for Twelve Departments, Identified by Number (49, 62, *etc.*).
Note: TC4 is the synthetic rate obtained after three iterations. Where the weight for a source has been eliminated or reduced, the value of the modulation coefficient (WM3) is shown.

fragmentary, owing to the difficulty of assembling, in 1993, management data for years past (1982, ...); in addition, the relationships used to draw an estimate of the net migration rate from each source were simplistic; and lastly, the method of synthesis was less elaborate.

It should be noted that the integration of other sources – income tax data in particular – can only further reinforce the effectiveness of the system.

Table 4
Mean of Distance in Retrospective Test

	TH	EDF	AF	EN	FE
1982	0.26	0.34	0.50	0.47	0.34
1983	0.28	0.33	0.48	0.47	0.32
1984	0.23	0.28	0.40	0.45	0.34
1985	0.24	0.31	0.48	0.44	0.32
1986	0.23	0.33	0.40	0.33	
1987	0.40	0.28	0.41	0.27	
1988	0.84	0.29	0.30	0.37	0.24
1989	0.97	0.21	0.30	0.33	0.35
Overall mean	0.43	0.30	0.41	0.39	0.32

Notes: -The number of rates per year is generally 96, except for FA (89) and EF (94).
-The "electoral file" source did not provide rates for 1986 or 1987.
-The "housing tax" source began to be distorted in 1987.
-The values of the differences correspond to rates expressed as a %.

8. SUPPLEMENTS

8.1 Sub-Departmental Levels

The use of some sources may become risky at a geographic level below the departmental level. There are various reasons for this: because the hypotheses on which the method is based become fragile, because the numbers are small, *etc.* This is especially the case with educational statistics.

However, it should be possible to operate the system for employment areas, or more specifically for cross-tabulations of department and employment area (there are approximately 420 such areas), which serve to ensure consistency with the departmental level. This should not involve too many risks, for the following reasons:

- a certain deterioration of performance in relation to the departmental estimates is acceptable, especially since the departmental estimates should be of good quality;
- the data from the income tax files should be quite useful;
- trend estimation and calibration on estimates at higher geographic levels (in this case the departmental estimates) both act as safeguards.

Of course, there is nothing prohibiting the use of the system to produce estimates for other sub-departmental geographic units.

At the departmental level, it does not seem useful to adapt the parameters (initial weights and norms) to population size; on the other hand, for sub-departmental

LAURENT, L., and GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE. Rennes.

LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E. Ottawa.

LAURENT, L., and GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE. Rennes.

LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E. Ottawa.

Survey Methodology, December 1990
Vol. 16, No. 2, pp. 217–237
Statistics Canada

Robust Small Area Estimation Combining Time Series and Cross-Sectional Data

D. PFEFFERMANN and L. BURCK[1]

ABSTRACT

The common approach to small area estimation is to exploit the cross-sectional relationships of the data in an attempt to borrow information from one small area to assist in the estimation in others. However, in the case of repeated surveys, further gains in efficiency can be secured by modelling the time series properties of the data as well. We illustrate the idea by considering regression models with time varying, cross-sectionally correlated coefficients. The use of past relationships to estimate current means raises the question of how to protect against model breakdowns. We propose a modification which guarantees that the model dependent predictors of aggregates of the small area means coincide with the corresponding survey estimators and we explore the statistical properties of the modification. The proposed procedure is applied to data on home sale prices used for the computation of housing price indexes.

KEY WORDS: Kalman filter; Linear constraints; State-space models.

1. INTRODUCTION

Statistical Bureaus are often confronted with the demand to provide reliable estimators for small area means. The problem with the production of such estimators is that the sample sizes within those areas are usually too small to allow the use of direct survey estimators. As a result, new estimators have been proposed in recent years which combine auxiliary information (obtained from a census or administrative records) with the survey data obtained from all the small areas. The common feature of these estimators is that they can be structured in general as a linear combination of two components: a "synthetic estimator" of the form $\bar{X}_i'\hat{\beta}$ where $\bar{X}_i$ represents the average auxiliary information at the small area level and $\hat{\beta}$ is a vector of estimated regression coefficients; and a "correction factor" of the form $(\bar{y}_i - \bar{x}_i\hat{\beta})$ where $\bar{y}_i$ and $\bar{x}_i$ are the sample means of the target and the auxiliary variables. The correction factors are used to account for the variability of the small area means not explained by the auxiliary variables. The major difference between the various estimators is in the approach followed to determine the weights assigned to the two components in the linear combination, ranging from a "design based approach" (Särndal and Hidiroglou 1989) to "empirical Bayes" (Fay and Herriot 1979) and "mixed linear models" (Battese, Harter and Fuller 1989, Pfeffermann and Barnard 1991).

Very few studies are reported in the literature on the possible use of the time series relationships of the data to further increase the efficiency of the small area estimators. This is despite the fact that many of the small area estimators are derived from repeated surveys such as labour force surveys. The econometric literature contains a vast number of studies on the combined modelling of time series and cross-sectional data, see *e.g.* Rosenberg (1973b), Johnson (1977, 1980), Maddala (1977, Chapter 7), Dielman (1983) and Pfeffermann and Smith (1985) for reviews. However, none of these studies is directed to the problem of estimating (predicting) small area means from survey data. Fitting time series models to survey data has been considered

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905. L. Burck, Unit for Statistical Analysis, Central Bureau of Statistics, Jerusalem 91130.

in the context of estimating aggregate population means, see the review papers of Smith (1979) and Binder and Hidiroglou (1988) and the more recent articles by Binder and Dick (1989), Tiller (1989) and Pfeffermann (1991). But again, these methods are not in routine use mainly because the classical survey estimators of the aggregate means are often almost as efficient when the models hold and more robust when the models fail to hold.

The situation is clearly different when dealing with a small area estimation problem; it seems to us that for this kind of problem, the use of time series models can be of great advantage. Although the exact nature of the model to be used in a particular application is obviously 'data dependent', the class of models we consider in the next section is broad enough to apply to many, if not most of the small area estimation problems arising in practice. These models have the further advantage that their estimation is relatively simple. Estimation issues are discussed in Section 3.

The use of a model always raises the question of how to protect against possible model failures and this question becomes even more sensitive when considering the use of a model for the production of official statistics. In Section 4 we consider this issue and propose a modification to the model dependent predictors which guarantees that for aggregates of the small area means for which the direct survey estimators can be trusted, the modified model predictors coincide with the survey estimators. The statistical properties of the modified predictors are explored. We conclude the article in Section 5 with empirical results which illustrate the performance of the model with and without the proposed modification. The data used for the illustrations are the sale prices of homes in the city of Jerusalem during the months of September 1985 through November 1989. These data are used routinely by the Central Bureau of Statistics in Israel for the computation of housing price indexes.

2. REGRESSION WITH CROSS-SECTIONALLY AND TIME VARYING COEFFICIENTS

2.1 A General Class of Models

In what follows we denote by $\underset{\sim}{Y}_{tk}$ the $n_{tk} \times 1$ vector of observations on a target variable Y, pertaining to an area k at time t, $k = 1, \ldots, K$, $t = 1, 2, \ldots$. We assume for convenience that $n_{tk} \geq 1$ but as becomes evident later on, the model permits that some of the areas not be observed at certain times. Let X_{tk} define the corresponding $n_{tk} \times (p + 1)$ design matrix of the auxiliary variables with a vector of ones as its first column. In many applications, the same row vector $\underset{\sim}{x}'_{tk}$ of auxiliary values applies to all the Y values of a given time so that $X_{tk} = \underset{\sim}{1}_{n_{tk}} \underset{\sim}{x}'_{tk}$ where $\underset{\sim}{1}_{n_{tk}}$ is a column vector of ones of length n_{tk}. This is the case when the only available data are the small area survey estimators. Confidentiality as well as processing costs often preclude the use of micro data on individual survey respondents. The theory described in this article is not restricted to the availability of the micro data (see the example in Section 2.2) but data availability has an obvious effect on model specifications and precision of estimation.

The regression model holding in area k at time t is defined as

$$\underset{\sim}{Y}_{tk} = X_{tk}\underset{\sim}{\beta}_{tk} + \underset{\sim}{\epsilon}_{tk};\ E(\underset{\sim}{\epsilon}_{tk}) = 0,\ E(\underset{\sim}{\epsilon}_{tk}\underset{\sim}{\epsilon}'_{tk}) = \sigma_k^2 I_{n_{tk}} \tag{2.1}$$

where $\underset{\sim}{\beta}'_{tk} = (\beta_{tk0}, \beta_{tk1}, \ldots, \beta_{tkp})$.

We define the (superpopulation) mean of the target variable values in area k at time t to be

$$\Theta_{tk} = E(M_{tk} \mid \underset{\sim}{\beta}_{tk}) = \underset{\sim}{\bar{X}}_{tk}\underset{\sim}{\beta}_{tk} \tag{2.2}$$

where

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} \quad \text{and} \quad \underset{\sim}{\bar{X}}_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} \underset{\sim}{x}'_{tki}$$

with $i = 1, \ldots, N_{tk}$ indexing the population units. Obviously, when $\underset{\sim}{x}'_{tki} \equiv \underset{\sim}{x}'_{tk}$, then $\underset{\sim}{\bar{X}}_{tk} = \underset{\sim}{x}'_{tk}$.

Let $\hat{\underset{\sim}{\beta}}_{tk}$ define an estimator for $\underset{\sim}{\beta}_{tk}$. Then $\hat{\Theta}_{tk} = \underset{\sim}{\bar{X}}_{tk} \hat{\underset{\sim}{\beta}}_{tk}$ and

$$\hat{M}_{tk} = \frac{1}{N_{tk}} \left[\sum_{i=1}^{n_{tk}} Y_{tki} + \sum_{i=n_{tk}+1}^{N_{tk}} \underset{\sim}{x}'_{tki} \hat{\underset{\sim}{\beta}}_{tk} \right] = \hat{\Theta}_{tk} + \frac{1}{N_{tk}} \left(\sum_{i=1}^{n_{tk}} (Y_{tki} - \underset{\sim}{x}'_{tki} \hat{\underset{\sim}{\beta}}_{tk}) \right)$$

implying that in the usual case of small sampling rates within the areas, $\hat{\Theta}_{tk}$ can also be considered as an estimator of the finite population mean M_{tk}. For this reason we no longer distinguish between the finite and superpopulation means.

The notable feature of (2.1) is that the coefficients $\underset{\sim}{\beta}_{tk}$ are allowed to vary both cross-sectionally and over time. The following equations specify the variation of the coefficients over time:

$$\begin{bmatrix} \beta_{tkj} \\ \beta_{kj} \end{bmatrix} = T_j \begin{bmatrix} \beta_{t-1,kj} \\ \beta_{kj} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{tkj}, \; j = 0, \ldots, p \tag{2.3}$$

where we use the notation β_{kj}, $j = 0, 1, \ldots, p$, to define fixed coefficients which we interpret below, and T_j to define fixed (2×2) matrices and where the residuals $\{\eta_{tkj}\}$ satisfy

$$E(\eta_{tkj}) = 0, \; E(\eta_{tkj}\eta_{tkt}) = \delta_{jt}, \; E(\eta_{tkj}\eta_{t-d,kt}) = 0 \quad \text{for} \quad d > 0. \tag{2.4}$$

The implication of (2.4) is that residuals of different coefficients pertaining to the same time t are allowed to be correlated but the serial and cross serial correlations are assumed to be zero.

Next, we illustrate the use of (2.3) by considering some simple cases:

(a) $T_j = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{kj} + \eta_{tkj}$ so that β_{kj} represents, in this case, a common mean. This is the well known Random Coefficient Regression Model (Swamy 1971) which is often used in econometric applications. Obviously, by postulating, $\text{var}(\eta_{tkj}) = 0$, the model reduces to the case of a fixed regression coefficient over time.

(b) $T_j = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$ which is the familiar random walk model, see *e.g.* Cooley and Prescott (1976) and LaMotte and McWhorter (1977) for application of this model in econometric studies. In this case the coefficient β_{kj} is redundant and should be omitted so that $T_j \equiv 1$.

(c) $T_j = \left[\begin{smallmatrix} \rho & 1-\rho \\ 0 & 1 \end{smallmatrix}\right]$ implies the first order autoregressive relationship $(\beta_{tkj} - \beta_{kj}) = \rho(\beta_{t-1,kj} - \beta_{kj}) + \eta_{tkj}$ considered by Rosenberg (1973a).

(d) $T_j = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \beta_{kj} + \eta_{tkj}$ which defines a local approximation to a linear trend (Kitagawa and Gersch 1984). The coefficient β_{kj} represents, in this case, a fixed slope.

It should be emphasized that different matrices T_j can be used for different coefficients β_{tkj}. In fact, by defining $\underset{\sim}{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \beta_{k1}, \ldots, \beta_{tkp}, \beta_{kp})$; $\tilde{T} = \text{diag}[T_0, T_1, \ldots, T_p]$, a block diagonal matrix with T_j as the j-th block; $\tilde{G} = I_{p+1} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ where I_{p+1} is the identity matrix of order $p + 1$ and $\otimes$ defines the Kronecker product and $\underset{\sim}{\eta}'_{tk} = (\eta_{tk0}, \eta_{tk1}, \ldots, \eta_{tkp})$, the combined model holding for the coefficients $\underset{\sim}{\beta}_{tk}$ can be written as

$$\underset{\sim}{\alpha}_{tk} = \tilde{T}\underset{\sim}{\alpha}_{t-1,k} + \tilde{G}\underset{\sim}{\eta}_{tk}; \quad E(\underset{\sim}{\eta}_{tk}) = \underset{\sim}{0}, E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{t-d,k}) = A_d\Delta \tag{2.5}$$

where $A_d = 1$ for $d = 0$ and $A_d = 0$ otherwise, and $\Delta = [\delta_{ij}]$ is defined by the variances and covariances δ_{ij} (equation 2.4).

The model defined by (2.5) specifies the variation of the regression coefficients of a specific area over time. The common approach to account for cross-sectional relationships between small area means is to allow for random small area effects which are time invariant $\{u_k\}$. The general model defined by (2.1) and (2.3) includes this case by writing $\underset{\sim}{Y}_{tk} = \underset{\sim}{1}_{n_{tk}}u_{tk} + X_{tk}\underset{\sim}{\beta}_{tk} + \underset{\sim}{\epsilon}_{tk} = X^*_{tk}\underset{\sim}{\beta}^*_{tk} + \underset{\sim}{\epsilon}_{tk}$, say, and specifying $u_{tk} = u_{t-1,k} + \eta_{tk}$ with $u_{ok} = 0$, $\text{var}(\eta_{1k}) = \sigma^2_\eta$ and $\text{var}(\eta_{tk}) = 0$ for $t > 1$ (compare with case (b) above). By assuming in addition the autoregressive relationship defined by case (c) for the intercept variable and fixing the other regression coefficients (case (a) with zero residual variances), the resulting model is similar to the model considered by Choudhry and Rao (1989) except that in their general formulation of the model the observation residuals of equation (2.1) are allowed to be serially correlated. Notice that equation (2.1) now contains two random "intercept terms" but the model is nonetheless identifiable. Choudhry and Rao assume that the only available data are the survey estimators so that the estimation of the serial correlations needs to be carried out externally, using the micro observations. Alternatively, a model accounting for the serial correlations can be postulated. Choudhry and Rao assume an AR(1) model in their study.

A more general way to account for the cross-sectional relationships between the small area means is to allow for non zero correlations between the residual terms η_{tkj} and η_{tmj} of the models specifying the time series variation of the regression coefficients β_{tkj} and β_{tmj} operating in areas k and m (equation 2.4). Often it is reasonable to assume that the correlations decay as the distance between the areas increases. This can be formulated as, $E(\eta_{tkj}, \eta_{tmj}) = \delta_{jj}\rho_j f_j(k,m)$, $k \neq m$, where $f_j(k,m)$ is a monotonic decreasing function of the distances $D(k,m)$. The case of geometrically decaying correlations is obtained by defining $f_j(k,m) = \rho_j^{|k-m|-1}$. The case of fixed correlations is obtained by specifying $f_j(k,m) \equiv 1$ and in what follows we consider this case only. Allowing for fixed cross-sectional correlations for all the regression coefficients can be formulated as

$$E(\underset{\sim}{\eta}_{tk}\underset{\sim}{\eta}'_{tm}) = D(\Delta)\emptyset, \quad k \neq m \tag{2.6}$$

where $D(\Delta)$ is the diagonal matrix with the variances δ_{jj} on the main diagonal and $\emptyset$ is another diagonal matrix composed of the correlations ρ_j.

Before concluding this section we present the model defined by (2.1), (2.5) and (2.6) in a state-space form. Presenting the model in this form has important computational advantages.

Let $\underset{\sim}{Y}'_t = (\underset{\sim}{Y}'_{t1}, \ldots, \underset{\sim}{Y}'_{tK})$ represent the vector of observations of length $n_t = \sum_k n_{tk}$ for all the areas at time t and let $\underset{\sim}{\epsilon}'_t = (\underset{\sim}{\epsilon}'_{t1}, \ldots, \underset{\sim}{\epsilon}'_{tK})$ represent the corresponding regression residuals. Define $Z_{tk} = [\underset{\sim}{1}_{ntk}, \underset{\sim}{0}_{ntk}, \underset{\sim}{x}_{tk1}, \underset{\sim}{0}_{ntk}, \ldots, \underset{\sim}{x}_{tkp}, \underset{\sim}{0}_{ntk}]$ where $\underset{\sim}{0}_{ntk}$ is a vector of zeroes of length n_{tk} and $\underset{\sim}{x}_{tkj}$ is the vector of values for the j-th auxiliary variable, $j = 1, \ldots, p$. Let Z_t be the block diagonal matrix composed of the matrices Z_{tk}. The matrix Z_t is of order $n_t \times [K \times 2 \times (p + 1)]$. Define also $\underset{\sim}{\alpha}'_t = (\underset{\sim}{\alpha}'_{t1}, \ldots, \underset{\sim}{\alpha}'_{tK})$, $\underset{\sim}{\eta}'_t = (\underset{\sim}{\eta}'_{t1}, \ldots, \underset{\sim}{\eta}'_{tK})$, $\Sigma_t = \text{Diag}[\sigma^2_1 \underset{\sim}{1}'_{nt1}, \ldots, \sigma^2_K \underset{\sim}{1}'_{ntK}]$, $T = I_K \otimes \tilde{T}$, and $G = I_K \otimes \tilde{G}$.

Using this notation, the model defined by (2.1), (2.5) and (2.6) can be written compactly as

$$\underset{\sim}{Y}_t = Z_t \underset{\sim}{\alpha}_t + \underset{\sim}{\epsilon}_t;\ E(\underset{\sim}{\epsilon}_t) = \underset{\sim}{0},\ E(\underset{\sim}{\epsilon}_t \underset{\sim}{\epsilon}_t') = \Sigma_t \tag{2.7}$$

$$\underset{\sim}{\alpha}_t = T\underset{\sim}{\alpha}_{t-1} + G\underset{\sim}{\eta}_t;\ E(\underset{\sim}{\eta}_t) = \underset{\sim}{0},\ E(\underset{\sim}{\eta}_t \underset{\sim}{\eta}_t') = \Lambda, \tag{2.8}$$

where $\Lambda = [\Lambda_{kl}]$, $k, \ell = 1, \ldots, K$ with $\Lambda_{kl} = \Delta$ when $k = \ell$ and $\Lambda_{kl} = D(\Delta)\emptyset$ when $k \neq \ell$. The matrices Λ_{kl} are $(p + 1) \times (p + 1)$.

The model defined by (2.7) and (2.8) conforms to the classical state-space formulation, see, *e.g.* Anderson and Moore (1979) and Harvey (1984). By this formulation, (2.7) is the observation equation and (2.8) is the state equation with $\underset{\sim}{\alpha}_t$ defining the state vector. The apparent advantage of restructuring the model in a state space form is that the vectors $\underset{\sim}{\alpha}_t$, and hence the population means Θ_{tk}, as well as the estimation error variances can be estimated conveniently by means of the Kalman filter. We discuss the use of the filter in sections 3 and 4.

2.2 Explicit Estimators of the Small Area Means

In order to illustrate how past and neighbouring data are used under the model to "strengthen" the small area estimators we consider the case where the same vector $\underset{\sim}{x}_{tk}$ of auxiliary values applies to all the units of a given area at a given time. In this case the observation equation can be formulated in terms of the sample means, *i.e.*

$$\bar{Y}_{tk} = \underset{\sim}{x}_{tk}' \underset{\sim}{\beta}_{tk} + \bar{\epsilon}_{tk};\ E(\bar{\epsilon}_{tk}) = 0,\ E(\bar{\epsilon}_{tk}^2) = \sigma_k^2/n_{tk},\ k = 1, \ldots, K. \tag{2.9}$$

Suppose that the regression coefficients follow a random walk (case (b) of equation 2.3) so that for area k

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj};\ E(\eta_{tkj}) = 0,\ E(\eta_{tkj}\eta_{tk\ell}) = \delta_{j\ell},\ j, \ell = 1, \ldots, p \tag{2.10}$$

and for areas $k \neq m$,

$$E(\eta_{tkj}\eta_{tmj}) = \delta_{jj}\rho_j;\quad E(\eta_{tkj}\eta_{tm\ell}) = 0,\ j \neq \ell. \tag{2.11}$$

The random walk model implies that the coefficients drift slowly away from their initial value with no inherent tendency to return to a mean value. Obviously, for residuals η_{tkj} such that $E(\eta_{tkj}^2) = 0$ the corresponding regression coefficients are fixed over time. Notice also that since $\underset{\sim}{\beta}_{tk} = \underset{\sim}{\beta}_{t-1,k} + \underset{\sim}{\eta}_{tk}$, the predictor of $\underset{\sim}{\beta}_{tk}$ at time $(t - 1)$ is the same as the predictor $\hat{\underset{\sim}{\beta}}_{t-1,k}$ of $\underset{\sim}{\beta}_{t-1,k}$.

Using the Kalman filter equations presented in section 3, it is shown in the Appendix that the estimator $\hat{\Theta}_{tk}$ of the small area mean Θ_{tk} (equation 2.2) can be structured in this case in the following form

$$\hat{\Theta}_{tk} = \underset{\sim}{x}_{tk}' \hat{\underset{\sim}{\beta}}_{t-1,k} + \left(1 - \frac{\sigma_k^2}{n_{tk}\nu_k^2}\right)\left(\bar{Y}_{tk} - \underset{\sim}{x}_{tk}' \hat{\underset{\sim}{\beta}}_{t-1,k}\right) + \frac{\sigma_k^2}{n_{tk}\nu_k^2} \sum_{\substack{m=1 \\ m \neq k}}^{K} \gamma_{km}\left(\bar{Y}_{tm} - \underset{\sim}{x}_{tm}' \hat{\underset{\sim}{\beta}}_{t-1,m}\right) \tag{2.12}$$

where the coefficients $\{\gamma_{km}\}$ are the partial regression coefficients in the regression of $e_{tk} = (\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$ against the prediction errors $\{e_{tm} = (\bar{Y}_{tm} - \underset{\sim}{x}'_{tm}\hat{\underset{\sim}{\beta}}_{t-1,m})\}$ obtained in the other areas and v_k^2 is the residual (unexplained) variance in the regression.

The estimator $\hat{\Theta}_{tk}$ is composed of three components: the "synthetic" estimator, $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$, where $\hat{\underset{\sim}{\beta}}_{t-1,k}$ is the optimal predictor of $\underset{\sim}{\beta}_{tk}$ based on all the observations up to and including time $t-1$, the "correction factor" $(\bar{Y}_{tk} - \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k})$ based on the prediction error in area k, and an "adjustment factor" based on the prediction errors observed for the other areas. The first two components correspond to the components of the classical small area estimators discussed in the introduction. Notice that the smaller the sample size n_{tk}, the smaller is the weight assigned to the current sample mean $\bar{Y}_{tk}$ in the estimation of Θ_{tk} and the larger is the weight assigned to the time series predictor $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$. The third component in the right hand side of (2.12) represents the information borrowed from neighbouring areas. The weight assigned to this component depends on the magnitude of the correlations ρ_j between the corresponding error terms $\{\eta_{tkj}\}$ in the models holding for the regression coefficients (equation 2.11). Obviously, when the regressions in the various areas are independent so that $\rho_j = 0$ for all j and hence $\gamma_{km} = 0$ for all m, the third component vanishes and the predictor $\hat{\Theta}_{tk}$ reduces to a weighted average of the current mean $\bar{Y}_{tk}$ and the time series predictor $\underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{t-1,k}$.

3. MODEL ESTIMATION AND INITIALIZATION USING THE KALMAN FILTER

3.1 Estimation of the Regression Coefficients by Means of the Kalman Filter

In this section we present the Kalman filter equations for the updating and smoothing of the state vectors $\underset{\sim}{\alpha}_t$ defined by the equations (2.7) and (2.8) (the area regression coefficients in our case). We assume that the V-C matrices $\sum_t$ and Λ are known. Estimation of these matrices is considered in section 3.2. The theory of the Kalman filter is developed in numerous publications (see *e.g.* Anderson and Moore 1979 and Meinhold and Singpurwalla 1983) and so we restrict the discussion to aspects most germane to the small area estimation problem.

Let $\hat{\underset{\sim}{\alpha}}_{t-1}$ be the best linear unbiased predictor (blup) of $\underset{\sim}{\alpha}_{t-1}$ based on all the data observed up to time $(t-1)$. Since $\hat{\underset{\sim}{\alpha}}_{t-1}$ is blup for $\underset{\sim}{\alpha}_{t-1}$, $\hat{\underset{\sim}{\alpha}}_{t|t-1} = T\hat{\underset{\sim}{\alpha}}_{t-1}$ is the blup of $\underset{\sim}{\alpha}_t$ at time $(t-1)$. Furthermore, if $P_{t-1} = E(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})(\hat{\underset{\sim}{\alpha}}_{t-1} - \underset{\sim}{\alpha}_{t-1})'$ is the V-C matrix of the prediction errors at time $(t-1)$, $P_{t|t-1} = TP_{t-1}T' + G\Lambda G'$ is the V-C matrix of the prediction errors $(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)$. (Follows straightforwardly from 2.8).

When a new vector of observations $[\underset{\sim}{Y}_t, Z_t]$ becomes available, the predictor of $\underset{\sim}{\alpha}_t$ and the V-C matrix P_{t-1} are updated according to the formulae

$$\begin{aligned} \hat{\underset{\sim}{\alpha}}_t &= \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1}Z_t'F_t^{-1}(\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1}) \\ P_t &= (I - P_{t|t-1}Z_t'F_t^{-1}Z_t)P_{t|t-1} \end{aligned} \tag{3.1}$$

where $\hat{\underset{\sim}{Y}}_{t|t-1} = Z_t\hat{\underset{\sim}{\alpha}}_{t|t-1}$ is the blup of $\underset{\sim}{Y}_t$ at time $(t-1)$ so that $\underset{\sim}{e}_t = (\underset{\sim}{Y}_t - \hat{\underset{\sim}{Y}}_{t|t-1})$ is the vector of innovations with V-C matrix $F_t = (Z_tP_{t|t-1}Z_t' + \sum_t)$.

The new data observed at time t can be used also for the updating (smoothing) of past estimators of the state vectors and hence for the updating of past estimators of the small area means. Denoting by t^* the most recent month with observations, the smoothing is carried out using the equations

$$\hat{\underline{\alpha}}_{t|t^*} = \hat{\underline{\alpha}}_t + P_t T' P_{t+1|t}^{-1}(\hat{\underline{\alpha}}_{t+1|t^*} - T\hat{\underline{\alpha}}_t)$$

$$P_{t|t^*} = P_t + P_t T' P_{t+1|t}^{-1}(P_{t+1|t^*} - P_{t+1|t})P_{t+1|t}^{-1} TP_t;\ t = 2, 3, \ldots, t^* \tag{3.2}$$

where $P_{t|t^*}$ is the V-C matrix of the prediction errors $(\hat{\underline{\alpha}}_{t|t^*} - \underline{\alpha}_t)$. Notice that $\hat{\underline{\alpha}}_{t^*|t^*} = \hat{\underline{\alpha}}_{t^*}$ and $P_{t^*|t^*} = P_{t^*}$ define the starting values for the smoothing equations.

Estimators of the small area means or aggregates of the means are obtained from the filtered (or smoothed) estimators of $\underline{\alpha}_t$ in a straightforward manner using the relationship $\hat{\Theta}_{tk} = \bar{X}_{tk}\hat{\underline{\beta}}_{tk} = \bar{\underline{Z}}'_{tk}\hat{\underline{\alpha}}_{tk} = \bar{\underline{Z}}'_{tk}A_{tk}\hat{\underline{\alpha}}_t$ where $\bar{\underline{Z}}'_{tk} = (1, 0, \bar{X}_{tk1}, 0, \ldots, \bar{X}_{tkp}, 0)$ and A_{tk} is the appropriate indicator matrix. Hence, if $\Theta_t^w = \sum_{k=1}^K w_k\Theta_{tk}$, then $\hat{\Theta}_t^w = \sum_{k=1}^K w_k\bar{\underline{Z}}'_{tk}A_{tk}\hat{\underline{\alpha}}_t = \underline{a}'_{tw}\hat{\underline{\alpha}}_t$, say. For given V-C matrices $\sum_t$ and Λ, the MSE's of the estimation errors are obtained as

$$E(\hat{\Theta}_{tk} - \Theta_{tk})^2 = \bar{\underline{Z}}'_{tk}A_{tk}P_tA'_{tk}\bar{\underline{Z}}_{tk} \quad \text{and} \quad E(\hat{\Theta}_{tk}^w - \Theta_{tk}^w) = \underline{a}'_{tw}P_t\underline{a}_{tw}. \tag{3.3}$$

Notice that the MSE's in (3.3) are with respect to the joint distribution of the observations $\{\underline{Y}_{tk}\}$ and the vectors of coefficients $\{\underline{\beta}_{tk}\}$ so that they represent average MSE's over the possible realizations of the area means.

3.2 Estimation of the V-C Matrices and Initialization of the Filter

The actual application of the Kalman filter requires the estimation of the unknown elements of the matrices $\sum_t$ and Λ and the initialization of the filter, that is, the estimation of the vector $\underline{\alpha}_o$ and the corresponding V-C matrix P_o of the estimation errors. In this section we describe simple estimation procedures which can be used for these purposes.

Assuming a normal distribution for the residual terms $\underline{\varepsilon}_t$ and $\underline{\eta}_t$ of equations (2.7) and (2.8), the log likelihood function of the vectors $\underline{Y}_{m+1}, \ldots, \underline{Y}_{t^*}$, conditional on the first m vectors $\underline{Y}_1, \ldots, \underline{Y}_m$, can be formulated as

$$L(\underline{\lambda}) = \text{constant} - \frac{1}{2}\sum_{t=m+1}^{t^*}(\log|F_t| + \underline{e}'_t F_t^{-1}\underline{e}_t) \tag{3.4}$$

where $\underline{\lambda}$ contains the unknown model variances and covariances written in a vector form. The scalar m defines the number of time periods needed to construct initial values for the Kalman filter. (For the random walk model considered in section 2.2, $m = 1$, provided that sufficient data are available in every area to allow the computation of the OLS estimators of the vectors of coefficients). The expression in (3.4) follows from the prediction error decomposition, see Schweppe (1965) and Harvey (1981) for details. For given matrices $\sum_t$ and Λ, the innovations $\underline{e}_t$ and the V-C matrices F_t can be obtained by application of the Kalman filter equations (3.1).

The computation of the likelihood function requires the initialization of the Kalman filter which can be carried out most conveniently by application of the approach proposed by Harvey and Phillips (1979). By this approach, the nonstationary components of the state vector are initialized with very large error variances which corresponds to postulating a noninformative prior distribution so that the corresponding state estimates can conveniently be taken as zeroes. (For the random walk model, initializing with a noninformative prior yields the OLS estimators after one time period, see Meinhold and Singpurwalla 1983, for a Bayesian formulation of the Kalman filter). The stationary components of the state vector are initialized by the corresponding unconditional means and variances which may be part of the unknown parameters defining the arguments of the likelihood function.

Maximization of the likelihood function (3.4) can be implemented using the method of scoring with a variable step length. In particular, let $\underset{\sim}{\lambda}_{(o)}$ define initial estimates of the unknown elements in $\underset{\sim}{\lambda}$. Then the method of scoring consists of solving iteratively the set of equations

$$\underset{\sim}{\lambda}_{(i)} = \underset{\sim}{\lambda}_{(i-1)} + r_i\{I[\underset{\sim}{\lambda}_{(i-1)}]\}^{-1} g[\underset{\sim}{\lambda}_{(i-1)}] \tag{3.5}$$

where $\underset{\sim}{\lambda}_{(i-1)}$ is the estimator of $\underset{\sim}{\lambda}$ as obtained in the $(i-1)$-th iteration, $I[\underset{\sim}{\lambda}_{(i-1)}]$ is the information matrix evaluated at $\underset{\sim}{\lambda}_{i-1}$ and $g[\underset{\sim}{\lambda}_{(i-1)}]$ is the gradient of the log likelihood evaluated at $\underset{\sim}{\lambda}_{i-1}$. The coefficient r_i is a variable step length introduced to guarantee that $L[\underset{\sim}{\lambda}_{(i)}] \geq L[\underset{\sim}{\lambda}_{(i-1)}]$ in every iteration. The value of r_i can be determined by a grid search procedure in the region $[0,1]$. The formulae for the k-th element of the gradient vector and the $k\ell$-th element of the information matrix are given in Watson and Engle (1983).

Having estimated the model variances and covariances, these estimates can be substituted for the true parameters in the Kalman filter equations (3.1) - (3.2) to yield the estimators of the regression coefficients and the V-C matrices and hence the small area estimators and their variances (see equation 3.3). Notice however that the estimated V-C matrices ignore the variability induced by the need to estimate the unknown elements contained in $\underset{\sim}{\lambda}$. Ansley and Kohn (1986) propose correction factors of order $1/t^*$ to account for this extra variation in state space modelling using first order Taylor approximations. Hamilton (1986) proposes a Monte Carlo procedure which consists of sampling from a multivariate normal distribution with mean given by the maximum likelihood estimator of the vector $\underset{\sim}{\lambda}$ and V-C matrix defined by the inverse of the information matrix, and estimating the state vectors for each random realization of the parameter values. This procedure is more flexible in terms of the assumptions involved and provides further insight into the sensitivity of the Kalman filter estimators to errors in the variance and covariance estimators. However, it is computationally more intensive.

4. MODIFICATIONS TO PROTECT AGAINST MODEL BREAKDOWNS

4.1 Description of the Problem and Proposed Modifications

The use of a model for small area estimation seems inevitable in view of the small sample sizes within the areas. However it raises the question of how to protect against model breakdowns. Testing the model every time that new data becomes available is often not practical, requiring instead the development of a "built-in mechanism" to ensure the robustness of the estimators when the model fails to hold.

One possibility is to modify the regression estimators derived in the various time periods so that they satisfy certain linear constraints obtained by equating aggregate means of the raw data with their expected fitted values under the model. More precisely, we propose to augment the model equation (2.1) by linear constraints of the form

$$\sum_k W_{tk}^{(\ell)} \sum_i Y_{tki} = \sum_k W_{tk}^{(\ell)} \sum_i \underset{\sim}{x}'_{tki}\,\underset{\sim}{\beta}_{tk} \quad \ell = 1, 2, \ldots, L(t),\ t = 1, \ldots, t^* \tag{4.1}$$

where the coefficients $W_{tk}^{(\ell)}$ are fixed, standardized weights such that $\sum_k n_{tk} W_{tk}^{(\ell)} = 1$. An example for such a constraint would be the equation

$$\sum_{k=1}^{K} N_{tk}\hat{M}_{tk} \Big/ \sum_{k=1}^{K} N_{tk} = \sum_{k=1}^{K} N_{tk}(\bar{x}'_{tk}\beta_{tk}) \Big/ \sum_{k=1}^{K} N_{tk} \tag{4.2}$$

where $\hat{M}_{tk}$ is the direct, survey estimator in area k. For $\bar{x}_{tk} \simeq \bar{X}_{tk}$, the equation (4.2) guarantees that the model dependent predictor of the aggregate population mean coincides with the corresponding survey estimator. Such a constraint can be justified by arguing that the survey estimators, although not reliable enough for estimating the small area means due to the small sample sizes, can be trusted when being combined for estimating the aggregate mean. Notice that "adding up" constraints are ordinarily imposed on statistical agencies anyway. Battese, Harter and Fuller (1988) and Pfeffermann and Barnard (1991) use a similar constraint for analysing cross-sectional surveys. Often, the small areas can be grouped into broader groups, with sufficient data in each of the groups to justify the use of the survey estimators for estimating the corresponding group means. In this case, one can impose several constraints of the form (4.2) where the summation is now over the areas belonging to the same group. Notice in this respect that in view of the correlations between the regression coefficients operating in the various areas, a constraint applied to a sub-set of the areas will modify the regression estimates in all the areas. We illustrate this property in the empirical study.

It is important to emphasize that the set of constraints in (4.1) does not represent external information about possible values of the regression coefficients. Rather, it serves as a "control system" to guarantee that the model estimators adjust themselves more rapidly to possible changes in the behavior of the regression coefficients. As a result, the variances of the modified regression estimators are slightly larger than the variances of the optimal estimators under the model. Obviously, when no such changes occur and the variances of the aggregate means are sufficiently small, one would expect the constraints to be satisfied approximately even without imposing them explicitly. As mentioned above, it is possible to incorporate several separate constraints in each time period but it is imperative that the variances of the corresponding aggregate means will be small enough to ensure that the modifications are indeed needed and do not interfere with the random fluctuation of the raw data.

4.2 Inference Incorporating the Linear Constraints

In Section 4.1 we proposed to amend the model equations (2.1) by imposing the set of constraints (4.1) thereby ensuring the robustness of the regression estimators against sudden drifts in the values of the coefficients.

Computationally, this can be implemented most conveniently by augmenting the vectors Y_t of equation (2.7) by the scalars $\sum_k W_{tk}^{(f)} \sum_i Y_{tki}$, augmenting the matrices Z_t by the corresponding row vectors $(W_{t1}^{(f)} 1'_{nt1} Z_{t1}, \ldots, W_{tK}^{(f)} 1'_{ntK} Z_{tK})$ and setting the respective variances of the residual terms to zero. The augmented set of equations, together with (2.8), form a pseudo state-space model which could be estimated using the Kalman filter equations (3.1). Notice that the pseudo V-C matrix $\Sigma_t^{(P)}$ of the augmented residual vector is no longer positive definite (the last $L(t)$ rows and columns of $\Sigma_t^{(P)}$ consist of zeroes) but this does not cause computational difficulties.

The drawback of applying the Kalman filter to the pseudo model is that the V-C matrices of the regression estimators fail to account for the actual variability of the aggregate means appearing in the left hand side of (4.1). In order to deal with this problem, we propose to amend the formula for the updating of the V-C matrix P_t (equation 3.1) so that the variances and covariances of the aggregate means will be taken into account.

Let $\underset{\sim}{Y}_t^{(A)}$ and $Z_t^{(A)}$ represent the augmented Y vector and Z matrix at time t and denote by $\Sigma_t^{(A)}$ the actual V-C matrix of the residual terms $[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t]$. The matrix $\Sigma_t^{(A)}$ is of order $[n_t + L(t)]$ with Σ_t in the first n_t rows and columns and the variances and covariances of the means $\Sigma_k W_{tk}^{(t)} \Sigma_i Y_{tki}$ among themselves and with the vector $\underset{\sim}{Y}_t$ in the remaining rows and columns. Denoting by $\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}$ the robust predictor of $\underset{\sim}{\alpha}_{t-1}$ as obtained at time $(t-1)$ using the pseudo model and by $P_{t-1}^{(A)}$ the actual V-C matrix of the errors $(\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_{t-1})$, the modified state estimator at time t is obtained as

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + P_{t|t-1}^{(A)} Z_t^{(A)'} \left(F_t^{(P)}\right)^{-1} \left[\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)} T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)}\right] \tag{4.3}$$

where $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)}T' + G\Lambda G')$ and $F_t^{(P)} = Z_t^{(A)} P_{t|t-1}^{(A)} Z_t^{(A)'} + \Sigma_t^{(P)}$ (Compare with 3.1). It is shown in the Appendix that the actual V-C matrix $P_t^{(A)}$ of the errors $(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t)$ satisfies the recursive equation

$$P_t^{(A)} = \left[I - K_t^{(P)} Z_t^{(A)}\right] P_{t|t-1}^{(A)} + K_t^{(P)} \left[\Sigma_t^{(A)} - \Sigma_t^{(P)}\right] K_t^{(P)}, \tag{4.4}$$

where $K_t^{(P)} = P_{t|t-1}^{(A)} Z_t^{(A)'} (F_t^{(P)})^{-1}$ is the pseudo Kalman gain. The first expression on the right hand side of (4.4) corresponds to the usual updating formula of the Kalman filter (compare with 3.1)). The second expression is a correction factor which accounts for the actual variances and covariances of the means $\Sigma_k W_{tk}^{(t)} \Sigma_i Y_{tki}$, not taken into account in the first expression.

The amended Kalman filter defined by the equations (4.3) and (4.4) produces robust predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ instead of the optimal, model dependent predictors, $\hat{\underset{\sim}{\alpha}}_t$ but otherwise uses the correct V-C matrices under the model. Thus, this filter can be used for the routine estimation of the vectors of coefficients and hence for the estimation of the small area means, and when the model holds it will give similar results to those obtained under the optimal filter. In periods where the model fails to hold, the updating formula (4.4) could be incorrect (depending on the particular model failures) but the predictors $\hat{\underset{\sim}{\alpha}}_t^{(A)}$ will nonetheless satisfy the linear constraints (4.1). The smoothing equations (3.2) can likewise be modified to satisfy the linear constraints.

5. EMPIRICAL RESULTS

5.1 Description of the Data and Model Fitted

In order to illustrate the important features of the class of models defined in Section 2, we fitted such a model to home sale prices in Jerusalem. The sale prices are recorded on a monthly basis and are routinely used by the Central Bureau of Statistics in Israel for the computation of monthly housing price indexes (HPI) adjusted for changes in quality. The HPI is computed separately for each city or group of cities and for each house size defined by the number of rooms, ranging from 1 to 5. The number of transactions carried out each month is very small in many of these cells and for 1 room apartments it occasionally happens that there are no transactions. The mean and standard deviation (S.D.) of the monthly number of transactions carried out during the period July 1987 - November 1989 are listed below.

Size	1	2	3	4	5
Mean	2.7	29.0	101.9	39.7	5.6
S.D.	2.6	12.9	50.4	18.8	3.5

The need to adjust for changes in quality results from the fact that the transactions performed are not under control, giving rise to large differences in quality from one month to the other particularly in the small cells. The following quality measure variables (QMV) are recorded for every transaction: $\tilde{X}^{(1)}$ - the apartment floor area, $\tilde{X}^{(2)}$ - the age of the apartment, $X^{(3)}$, $X^{(4)}$ - dummy variables defining districts within the city.

The problems involved in the computation of the HPI and the method used in Israel are discussed at length in a recent article by Pfeffermann, Burck and Ben-Tuvia (1989). The following model was proposed by the authors as an alternative to the model in current use. The triple index "tki" defines the i-th transaction of size k in month t with Y_{tki} standing for the log of the sale price and $X_{tki}^{(j)} = \log(\tilde{X}_{tki}^{(j)})$, $j = 1, 2$.

$$Y_{tki} = \beta_{tk0} + \beta_{tk1}X_{tki}^{(1)} + \beta_{tk2}X_{tki}^{(2)} + \beta_{tk3}X_{tki}^{(3)} + \beta_{tk4}X_{tki}^{(4)} + \epsilon_{tki} \tag{5.1}$$

$$\begin{aligned} \beta_{tk0} &= \beta_{t-1,k0} + \beta_{k0} + \eta_{tk0} \\ \beta_{tkj} &= \beta_{t-1,kj} + \eta_{tkj},\ j = 1, \ldots, 4, \end{aligned} \tag{5.2}$$

with the error terms ϵ_{tki} and η_{tkj} satisfying the assumptions (2.1), (2.4) and (2.5). Notice that the model assumed for the intercept term is the local approximation to a linear trend defined under case (d) of Section (2.1). The model assumed for the other coefficients is the random walk model defined under case (b).

The regression defined by (5.1) forms the basis for the construction of an HPI adjusted for changes in quality. By fixing the values of the QMV's at their average population values which are constant over time, (the values of these variables are adjusted approximately every five years), average sale prices can be computed using (5.1) and these averages are comparable between months since they refer to homes of similar qualities.

Pfeffermann, Burck and Ben-Tuvia discuss the considerations in selecting the model defined by (5.2) for the regression coefficients. They show empirical results which validate the fitness of the model. However, the results of that study were obtained by fitting the model to each cell separately, that is, without accounting for the cross-sectional relationships of the regression coefficients. This aspect of the model is explored in the present study. Another major purpose of the empirical study is to illustrate the performance of the modifications proposed in Section 4 to protect against model breakdowns.

5.2 Estimation of the Model

The model defined by (5.1) and (5.2) can be put in a state-space form similar to (2.7) and (2.8). In fact, the vectors $\underline{\alpha}_t$ and the matrices Z_t, T and G assume, in this case, simple structures, since for $j = 1, \ldots, 4$, $\beta_{kj} \equiv 0$ (see case (b) of Section 2.1). Thus, $\underline{\alpha}'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \ldots, \beta_{tk4})$, $Z_{tk} = [\underline{1}_{ntk}, \underline{0}_{ntk}, \underline{X}_{tk}^{(1)}, \ldots, \underline{X}_{tk}^{(4)}]$, $\tilde{T} = [\underline{e}_1, \underline{e}_1 + \underline{e}_2, \underline{e}_3, \ldots, \underline{e}_6]$, a 6×6 matrix with $\underline{e}_j$ having a one in position j and zeroes elsewhere and $\tilde{G} = [\underline{e}_1, \underline{e}_3, \ldots, \underline{e}_6]$ which is 6×5. The matrix Δ is defined as in (2.5). The vector $\underline{\alpha}_t$ and the matrices Z_t, T, G and Λ are obtained from the vectors $\{\underline{\alpha}_{tk}\}$ and the matrices $\{Z_{tk}\}$, $\tilde{T}$, $\tilde{G}$ and Δ in the same way as in (2.7) and (2.8).

Having set the model in a state-space form we next attempted to estimate the unknown variances and covariances using the method of scoring algorithm described in Section 3.2. As it turned out, however, the computer time needed for convergence was way beyond the capacity of the IBM 1481 mainframe used for this study. Notice that the number of unknown parameters of the combined state-space model is $\dim(\underline{\lambda}) = 25$ whereas the dimension of the

state vectors and hence the dimension of the corresponding V-C matrices is $\dim(\alpha_t) = 30$. The total number of observations per month ranges from 55 to 353. The computer program written for this study uses numerical derivatives so that each iteration of the method of scoring requires a separate sweep through all the data with each sweep involving $[\dim(\lambda) + 1]$ computations of the state vector $\hat{\alpha}_t$ and the V-C matrix P_t (equation 3.1) at each point in time. These computations are needed in order to evaluate the log likelihood functions and hence the corresponding derivatives. It is clear therefore that the computational costs increase with the length of the series, the number of observations, the size of the state vector and the number of unknown parameters.

In order to deal with this problem we estimated the variance σ_k^2 (equation 2.1) and the matrix Δ (equation 2.5) separately for each of the five apartment sizes using the time series of observations corresponding to each size and then estimated the correlations ρ_j (equation 2.6) by a crude, grid search procedure. We found that setting $\rho_j = \frac{1}{2}$ for every j gives satisfactory results both in terms of the behaviour of the innovations (the one step ahead prediction errors) and in terms of the smoothness of the regression coefficients corresponding to apartments of size one and five where the monthly sample sizes are very small. Notice that by estimating the variances and covariances defining the time series relationships of the regression coefficients separately for each size, one is more flexible in terms of the model assumptions although there is some loss of efficiency if the variances and covariances are indeed the same across the different sizes.

5.3 Results

Pfeffermann, Burck and Ben-Tuvia (1989) illustrate the adequacy of the time series models fitted to the various apartment sizes. As mentioned earlier, our purpose in this study is to compare the results obtained with and without the accounting for the cross-sectional correlations and to illustrate the performance of the modifications (4.1) in protecting against model breakdowns.

In order to sharpen the comparisons as much as possible, we deliberately inflated the Y-values by 5 percent in each of the following four months: October 1987, November 1988, January 1989 and May 1989. Thus all the Y-values of all the apartment sizes corresponding to the months October 1987 - October 1988 were inflated by 5 percent, the Y-values corresponding to November 1988 - December 1988 were inflated by 10.25 percent (5 percent on top of the previous 5 percent) and so forth. These kinds of model breakdowns (although obviously not in such magnitudes) may result from intentional devaluations of the currency and are of main concern when modeling sale prices. See Pfeffermann, Burck and Ben-Tuvia for further discussion. Similar model breakdowns may occur, for example, with series of unemployment rates in periods of abrupt economic recessions.

Table 1 shows the average mean squared errors (AMSE) of the model residuals $\hat{\epsilon}_{tki} = (Y_{tki} - \hat{\beta}_{tko} - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{tkj})$ and the model innovations $e_{tki} = [Y_{tki} - (\hat{\beta}_{t-1,k0} + \beta_{k0}) - \sum_{j=1}^{4} X_{tki}^{(j)} \hat{\beta}_{t-1,kj}]$ (see equations 5.1 and 5.2), separately for each of the five apartment sizes. The AMSE's were computed as $\text{AMSE}_k(\epsilon) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} \hat{\epsilon}_{tki}^2)$; $\text{AMSE}_k(e) = 1/N \sum_{t=1}^{N} (1/n_t \sum_{i=1}^{n_t} e_{tki}^2)$ where $t = 1, \ldots, N$ indexes the months of July 1987 - November 1989. We distinguish between four different estimators of the regression coefficients as defined by whether the model accounts for the cross-sectional correlations ($\rho_j \equiv \frac{1}{2}$), ($\rho_j \equiv 0$) and by whether or not the estimators are modified to protect against the model breakdowns (abbreviated as "Rob. Inc." and "No Rob." in the table). The modifications were carried out by augmenting the observation equation of each month by three linear constraints of the form 4.2. These constraints forced the aggregate means of the fitted values in each of the three

Table 1

Average Mean Squared Errors of Residuals and Innovations With and Without the Accounting for Cross-sectional Correlations and the Inclusion of the Robustness Modifications, by Size

Apt. Size	Mean Squared Errors of Innovations				Mean Squared Errors of Residuals			
	$\rho \equiv ½$		$\rho \equiv 0$		$\rho \equiv ½$		$\rho \equiv 0$	
	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.
1	.141	.134	.176	.218	.021	.027	.056	.092
2	.070	.090	.084	.123	.021	.039	.023	.070
3	.065	.090	.070	.197	.017	.042	.019	.143
4	.067	.123	.072	.198	.019	.066	.021	.141
5	.067	.114	.077	.193	.023	.033	.065	.106

districts to coincide with the corresponding means of the observed values. When incorporating the constraints, the model was fitted using the amended Kalman filter as defined by the equations (4.3) and (4.4).

In order to illustrate the performance of the four sets of regression estimators in the various months and in particular, in and around the months where we inflated the data, we plotted the monthly MSE's of the innovations and residuals as obtained for 3 and 5 room apartments. The plots are shown in Figures 1 to 4. Notice that the values of Table 1 for 3 and 5 room apartments are correspondingly the averages of the values shown in the four figures.

The main conclusions from the table and the graphs are as follows:

Accounting for the cross-sectional correlations and including the linear constraints to protect against the model breakdowns yields better results than in the other cases considered. This outcome is most prominent in the cells of 1 and 5 room apartments where the sample sizes in each month are very small. In the other three cells, there are only small differences between the case ($\rho \equiv ½$, Rob. Inc.) and the case ($\rho \equiv 0$, Rob. Inc.) which could be expected since as the number of observations in each month increases, there is less borrowing of information from neighbouring cells (small areas in the more general context). The situation is different, however, when the linear constraints are removed. Accounting for the cross-sectional correlations yields in this case much better results than when not accounting for them and this is true for all the apartment sizes. Thus, by borrowing information from one cell to the other, the estimators of the regression coefficients adapt themselves much more rapidly to the sudden drifts in the data as seen also more directly in the figures [The four peaks in each graph are in the months where the data were inflated and as can be seen, the graphs corresponding to the case ($\rho \equiv ½$, No Rob.) return to their normal level of the months before the inflation much faster than the graphs representing the case ($\rho \equiv 0$, No Rob.)

Another interesting comparison is between the case where the linear constraints are included and the case where they are not. Clearly, the inclusion of the constraints improves the results substantially when accounting for the serial correlations and the improvements are even more prominent when the serial correlations are set to zero. It is interesting to compare in this context the figures exhibiting the monthly MSE's of the innovations with the figures exhibiting the monthly MSE's of the residuals. In the four months where we inflated the data the MSE's of the innovations are high which is obvious since the innovations are the differences between the observations and their predictors from previous months. Still, when the linear constraints are included, the MSE's return to their normal level right after the months of inflation. As

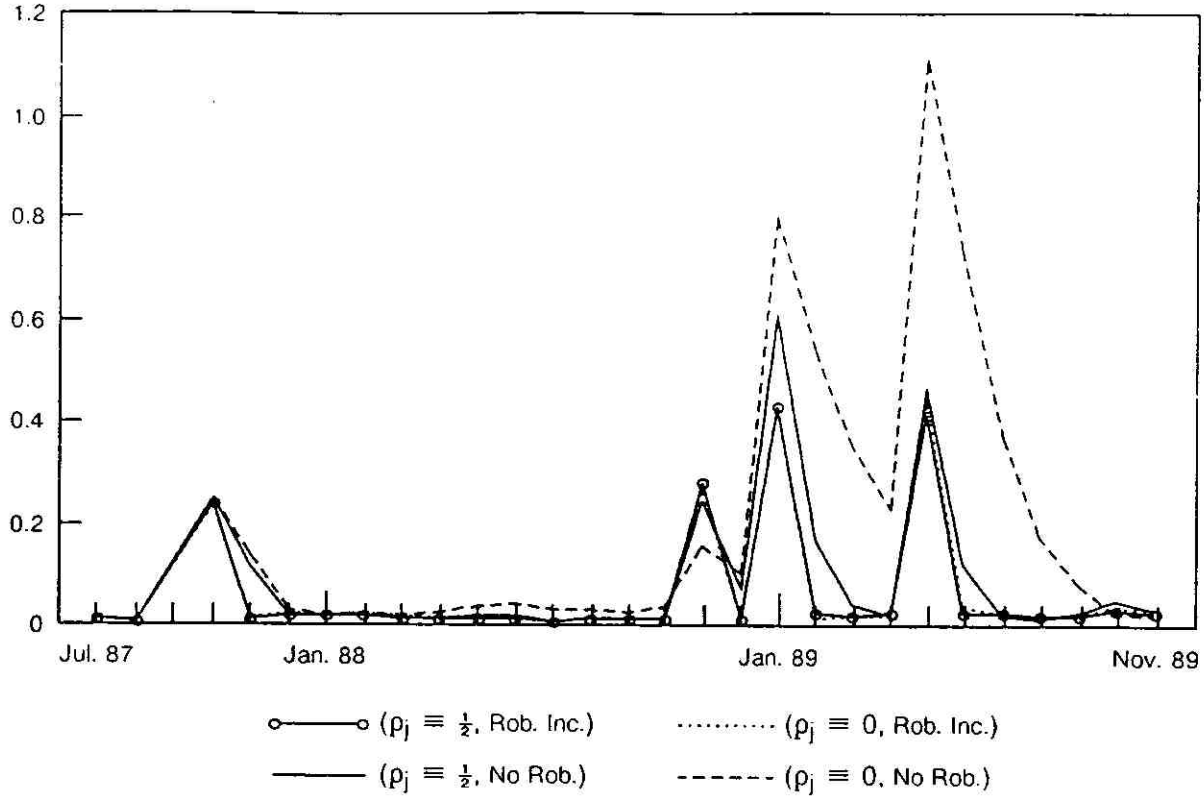

Figure 1 Monthly Mean Squared Errors of Innovations, 3 Room Apartments

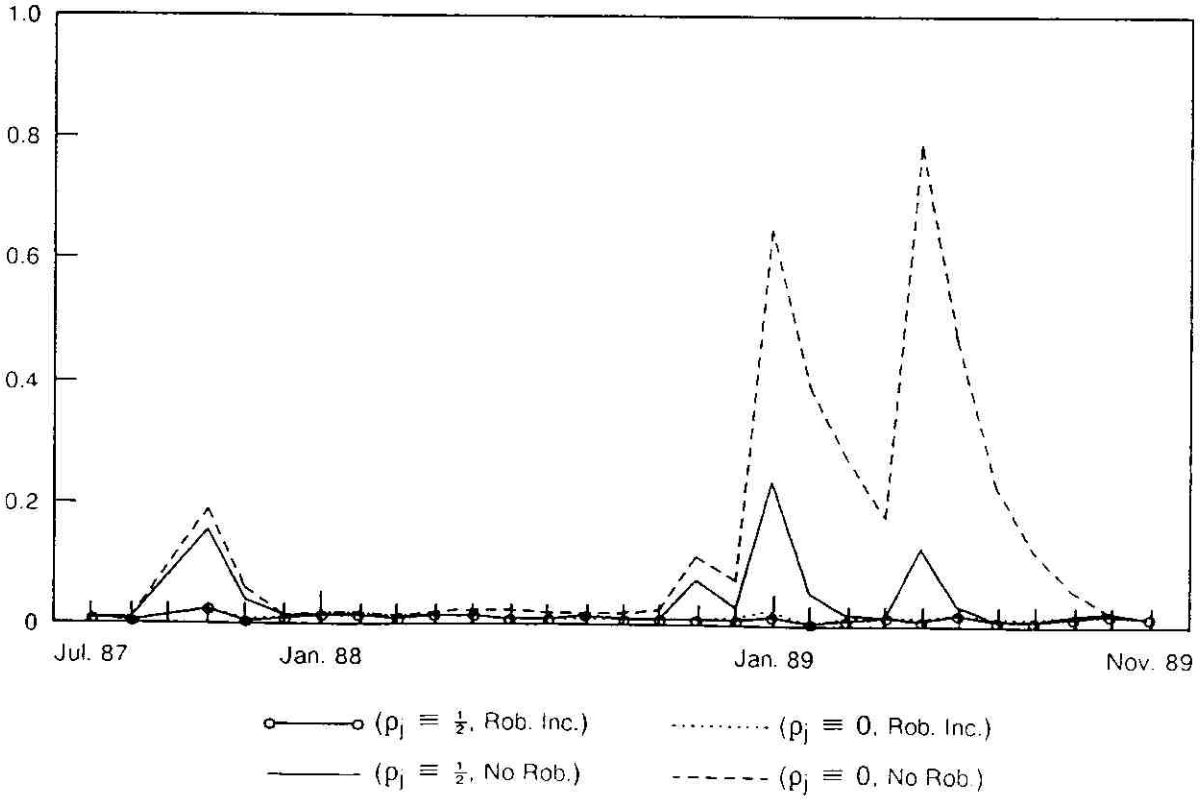

Figure 2 Monthly Mean Squared Errors of Residuals, 3 Room Apartments

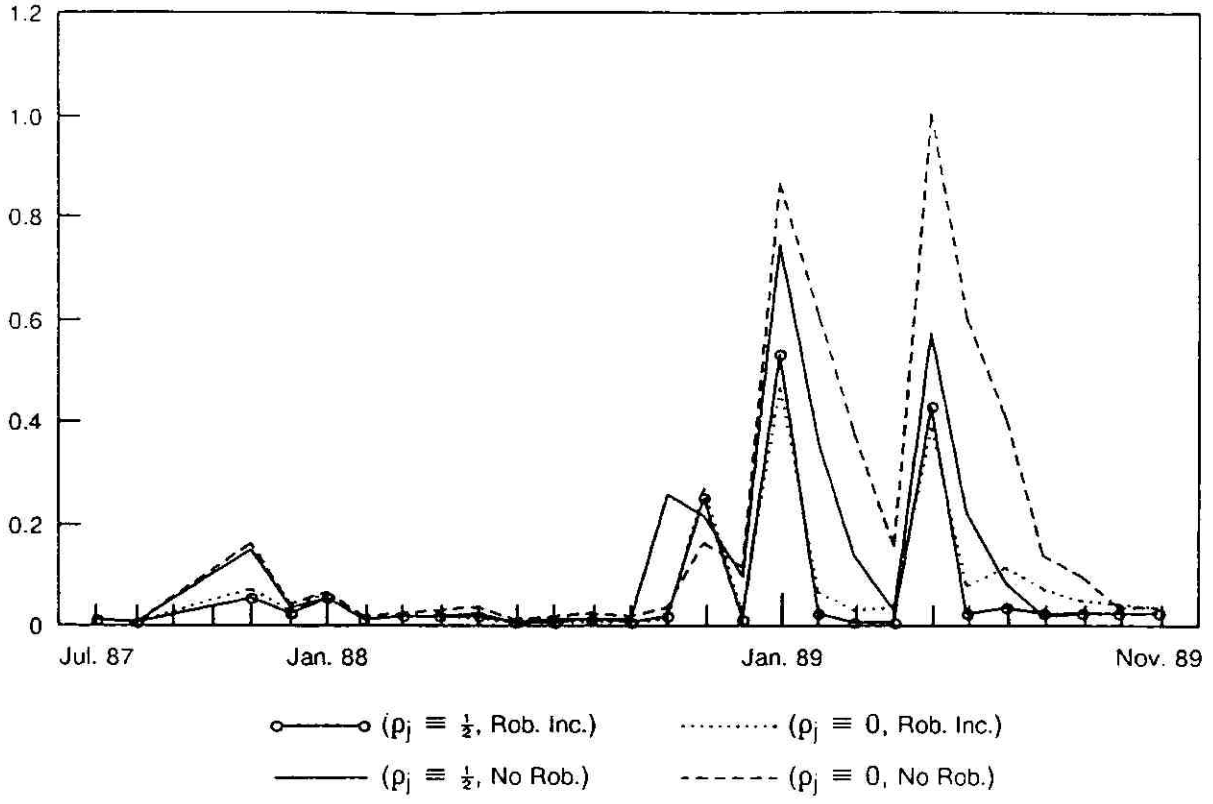

Figure 3 Monthly Mean Squared Errors of Innovations, 5 Room Apartments

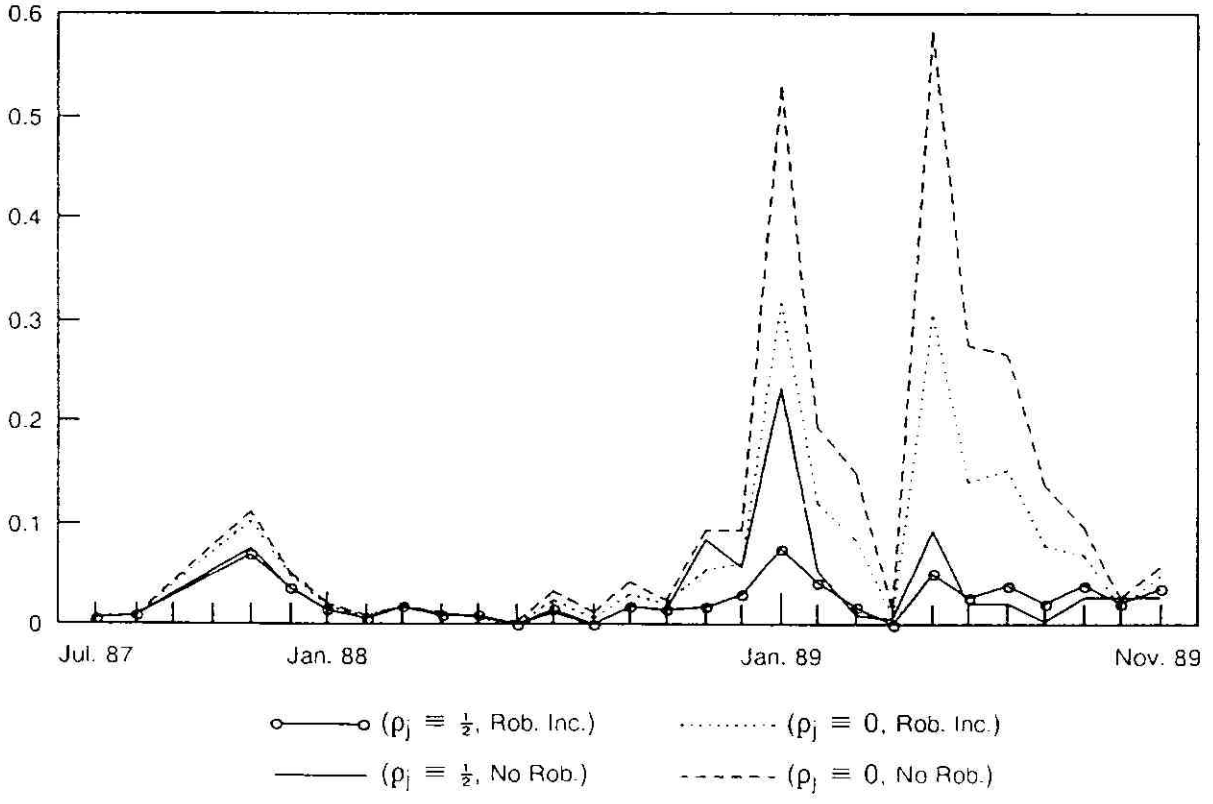

Figure 4 Monthly Mean Squared Errors of Residuals, 5 Room Apartments

for the residuals, once the linear constraints are included, there is practically no increase in the MSE values in the months of inflation in the case of 3 room apartments and, when accounting for the serial correlations, only a slight increase in the case of 5 room apartments. However, when ignoring the serial correlations, the residual MSE's for 5 room apartments are much larger in the months of inflation than in the other months even when imposing the constraints. This outcome has a simple explanation. The linear constraints are imposed on the aggregate means of the fitted values in each district but since the number of observations in 5 room apartments is a small fraction of the total number of observations, the constraints alone have a relatively small effect on the estimated regression coefficients in this cell. On the other hand, the constraints have a large effect on the estimated coefficients in the other cells so that when accounting for the cross-sectional correlations, the estimators corresponding to 5 room apartments are also modified since they are correlated with the other coefficients.

The way by which the linear constraints protect against sudden drifts in the data is illuminated in Figure 5 where we plotted the monthly intercept estimates for 3 room apartments.

As can be seen, with the linear constraints included, the intercept adapts itself to the new level of the data in the same month that the inflation occurs. Without the inclusion of the constraints, the adaption to the new level of the data takes several months. The plot of the monthly intercept estimates of 5 room apartments does not have this nice pattern since with the small sample sizes observed each month, the effect of the inflation is to alter also the other regression coefficients.

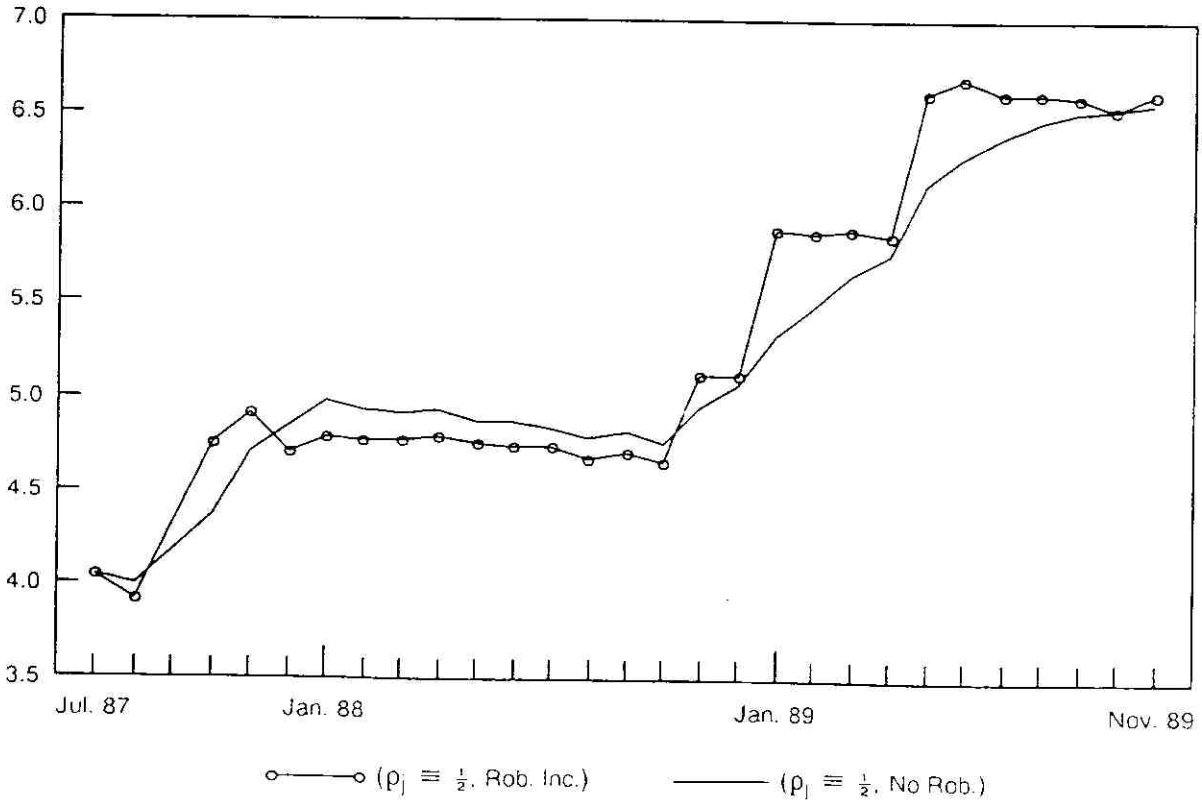

Figure 5 Monthly Estimates of Intercept, 3 Room Apartments

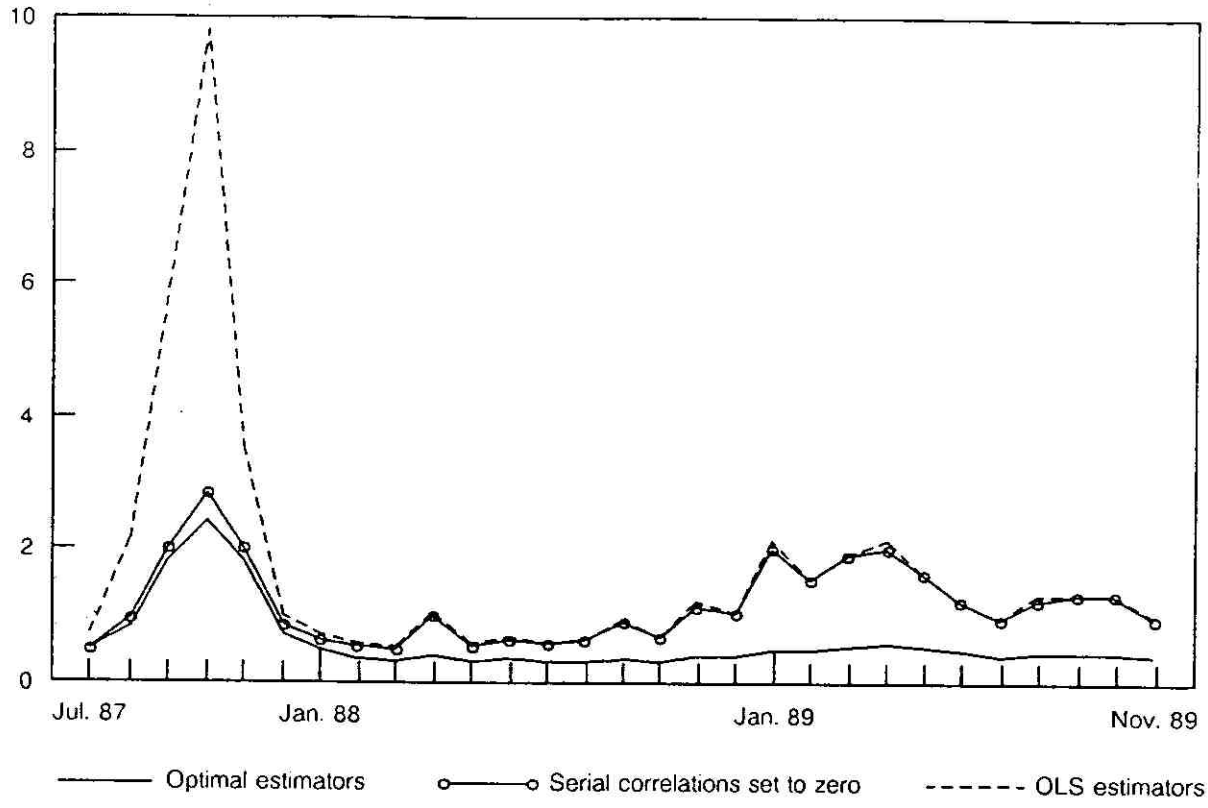

Figure 6 Variances of Estimators of Cell Means ($\times 10^4$), 3 Room Apartments

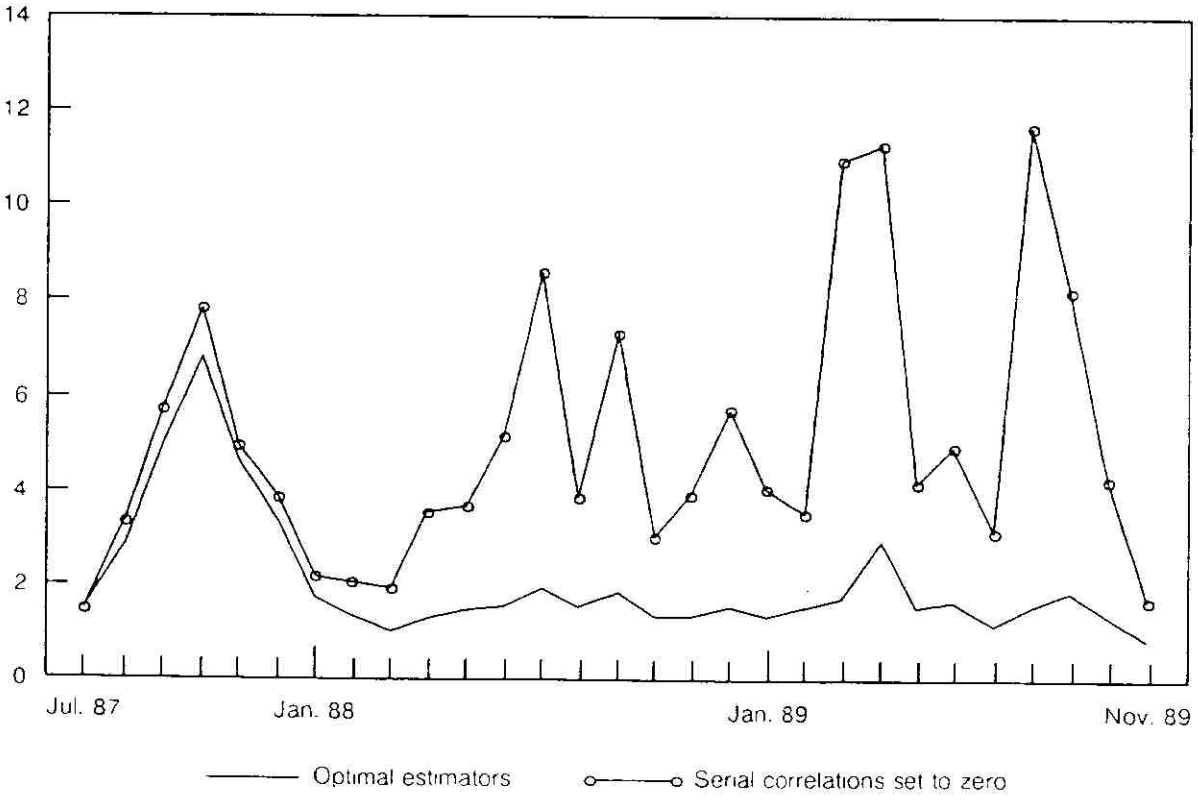

Figure 7 Variances of Estimators of Cell Means ($\times 10^4$), 5 Room Apartments

Our discussion so far centered on the empirical distribution of the model residuals and innovations. A major application of small area estimation is the prediction of the small area means (equation 2.2). Clearly, when a model yields residuals with well behaved properties it can also be expected to yield good estimators for the population means. Nevertheless, it is interesting to compare the theoretical variances of the small area means estimators as obtained with and without the accounting for the cross-sectional correlations, under the model which accounts for these correlations with $\rho_j \equiv \frac{1}{2}$. This comparison permits the assessment of the loss in efficiency when the serial correlations are ignored.

Figures 6 and 7 show the monthly variances of the cell mean estimators as obtained for 3 and 5 room apartments. (The variances have been multiplied by 10^4.) The figure for 3 room apartments also contains the variances of the ordinary least squares (OLS) estimators of the population means, that is, the variances of the estimators when estimating the regression coefficients in each month by OLS. These estimators are not operational in the case of 5 room apartments because of the very small monthly sample sizes.

The important conclusion drawn from the two figures is that by accounting for the cross-sectional correlations the variances of the resulting estimators can be reduced quite substantially, depending on the sample sizes. This is obviously the case in the case of 5 room apartments but is also true for 3 room apartments despite the fact that the sample sizes in these cells are relatively very large. The large sample sizes ordinarily obtained for 3 room apartments make the OLS estimators quite comparable to the estimators obtained when ignoring the cross-correlations in the estimation of the population means. Notice however the big gap between the variance of the OLS estimator and the variance of the other two estimators in October 1987. In this month there were only 10 observations of 3 room apartments and it is here where the use of the past data has its main impact even when ignoring the cross-sectional correlations. (The number of observations for 3 room apartments in November 1987 is 28; in all the other months there are at least 46 observations.)

Another important outcome arising from the two figures is the much greater stability of the variances of the optimal estimators under the model as compared to the variances of the estimators which ignore the cross-sectional correlations. Notice in this respect that the differences in the variances from one month to the other depend not only on the sample sizes in each month but also on the values of the explanatory variables (the design matrix) and the amount of past data observed. Still, it is the sample sizes which mostly explains the differences in the variances of the estimators particularly towards the end of the series.

ACKNOWLEDGEMENT

This article was written while the first author was on sabbatical leave at Statistics Canada under its Research Fellowship program. The authors would like to thank a referee for helpful comments.

APPENDIX

a) Derivation of Equation (2.12)

When $\underset{\sim}{x}_{tki} = \underset{\sim}{x}_{tk}$, $\hat{\Theta}_{tk} = \underset{\sim}{x}'_{tk}\hat{\underset{\sim}{\beta}}_{tk} = \underset{\sim}{z}'_{tk}\hat{\underset{\sim}{\alpha}}_{tk}$ so that $\hat{\underset{\sim}{\Theta}}_t = (\hat{\Theta}_{t1}, \ldots, \hat{\Theta}_{tK})' = Z_t\hat{\underset{\sim}{\alpha}}_t$. Also, for the random walk model the matrix T is the identity matrix and by equation (3.1)

$$Z_t\hat{\underset{\sim}{\alpha}}_t = Z_t\hat{\underset{\sim}{\alpha}}_{t-1} + \left(Z_tP_{t|t-1}Z_t'\right)F_t^{-1}\left(\underset{\sim}{Y}_t - Z_t\hat{\underset{\sim}{\alpha}}_{t-1}\right) =$$

$$\left(I - \Sigma_tF_t^{-1}\right)\underset{\sim}{Y}_t + \Sigma_tF_t^{-1}Z_t\hat{\underset{\sim}{\alpha}}_{t-1} \tag{A1}$$

since $F_t = (Z_tP_{t|t-1}Z_t' + \Sigma_t)$. Suppose for convenience that $k = 1$ and define

$$F_t = \begin{bmatrix} f_{11}, \underset{\sim}{f}_1' \\ \underset{\sim}{f}_1, F_{22} \end{bmatrix} \quad \text{and} \quad H_t = F_t^{-1} = \begin{bmatrix} h_{11}, \underset{\sim}{h}_1' \\ \underset{\sim}{h}_1, H_{22} \end{bmatrix} \quad \text{were} \quad f_{11} \quad \text{and} \quad h_{11}$$

are scalars, $\underset{\sim}{f}_1'$ and $\underset{\sim}{h}_1'$ are $[1 \times (K-1)]$ and F_{22} and H_{22} are $[(K-1) \times (K-1)]$. Using this notation, it follows from (A1) that

$$\hat{\Theta}_{t1} = \left(1 - \frac{\sigma_1^2}{n_{t1}}h_{11}\right)\bar{Y}_{t1} + \frac{\sigma_1^2}{n_{t1}}h_{11}\left(\underset{\sim}{x}_{t1}'\hat{\underset{\sim}{\beta}}_{t-1,1}\right) - \frac{\sigma_1^2}{n_{t1}}\sum_{k=2}^{K} h_{11}\frac{h_{1k}}{h_{11}}\bar{e}_{tk}. \tag{A2}$$

Let $\underset{\sim}{\gamma}_1' = (\gamma_{12}, \ldots, \gamma_{1K}) = \underset{\sim}{f}_1'F_{22}^{-1}$ defines the partial regression coefficients in the regression of $\bar{e}_{t1}$ on $(\bar{e}_{t2}, \ldots, \bar{e}_{tK})$ and $v_1^2 = (f_{11} - \underset{\sim}{f}_1'F_{22}^{-1}\underset{\sim}{f}_1)$ define the residual variance in the regression.

Equation (2.12) follows directly from (A2) since

$$\underset{\sim}{f}_1'F_{22}^{-1} = -\frac{1}{h_{11}}\underset{\sim}{h}_1'; \quad \left(f_{11} - \underset{\sim}{f}_1'F_{22}^{-1}\underset{\sim}{f}_1\right)^{-1} = h_{11} \tag{A3}$$

by well known properties of the inverse of a partitioned matrix.

b) Derivation of Equation (4.4)

By (4.3),

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} = \left(I - K_t^{(P)}Z_t^{(A)}\right)T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} + K_t^{(P)}\underset{\sim}{Y}_t^{(A)}. \tag{A4}$$

Hence,

$$\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t = \left(I - K_t^{(P)}Z_t^{(A)}\right)\left(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t\right) + K_t^{(P)}\left(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t\right). \tag{A5}$$

The prediction errors $\left(T\hat{\underset{\sim}{\alpha}}_{t-1}^{(A)} - \underset{\sim}{\alpha}_t\right)$ are independent of the residuals $\left(\underset{\sim}{Y}_t^{(A)} - Z_t^{(A)}\underset{\sim}{\alpha}_t\right)$ and so,

$$P_t^{(A)} = E\left[\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)\left(\hat{\underset{\sim}{\alpha}}_t^{(A)} - \underset{\sim}{\alpha}_t\right)'\right] = Q_tP_{t|t-1}^{(A)}Q_t' + K_t^{(P)}\Sigma_t^{(A)}K_t^{(P)\prime}, \tag{A6}$$

where we denote for convenience $Q_t = (I - K_t^{(P)}Z_t^{(A)})$.

By definition of the matrix $F_t^{(P)}$ (see below 4.3), equation (A6) can be written in the form

$$P_t^{(A)} = Q_t P_{t|t-1}^{(A)} - P_{t|t-1}^{(A)} Z_t^{(A)'} K_t^{(P)'} + K_t^{(P)} F_t^{(P)} K_t^{(P)'}$$

$$+ K_t^{(P)} \left(\Sigma_t^{(A)} - \Sigma_t^{(P)} \right) K_t^{(P)'} \qquad \text{(A7)}$$

which implies the relationship (4.4) by straightforward algebra.

REFERENCES

ANDERSON, B.O.D., and MOORE, J.B. (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

ANSLEY, C.F., and KOHN, R. (1986). Prediction mean squared error for State Space models with estimated parameters. *Biometrika*, 73, 467-473.

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.

BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, Vol. 6, (Eds.), P.R. Krishnaiah and C.R. Rao, Amsterdam: Elsevier Science, 187-211.

CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Analysis of Data in Time*. Ottawa: Statistics Canada (to appear).

COOLEY, T.F., and PRESCOTT, E.C. (1976). Estimation in the presence of stochastic parameter variation. *Economitrica*, 44, 167-184.

DIELMAN, T.E. (1983). Pooled cross-sectional and time series data: A survey of current statistical methodology. *The American Statistician*, 37, 111-122.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a State-Space model. *Journal of Econometrics*, 33, 388-397.

HARVEY, A.C. (1981). *Time Series Models*. Deddington, Oxford: Philip Allan.

HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

KITAGAWA, G., and GERSCH, W. (1984). A smoothness priors State-Space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79, 378-389.

JOHNSON, L.W. (1977). Stochastic parameter regressions: An annotated bibliography. *International Statistical Review*, 45, 257-272.

JOHNSON, L.W. (1980). Stochastic parameter regression: An additional annotated bibliography. *International Statistical Review*, 48, 95-102.

LaMOTTE, L.R., and McWHORTER, A. (1977). Estimation, testing and forecasting with random coefficient regression models. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 814-817.

MADDALA, G.S. (1977). *Econometrics*. Kogakusta: McGraw-Hill.

MEINHOLD, R.J., and SINGPURWALLA, N.D. (1983). Understanding the Kalman filter. *The American Statistician*, 37, 123-127.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 163-175.

PFEFFERMANN, D., and BARNARD, C. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* (to appear).

PFEFFERMANN, D., BURCK, L., and BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Analyisis of Data in Time*. Ottawa: Statistics Canada.

PFEFFERMANN, D., and SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.

ROSENBERG, B. (1973a). The analysis of cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, 2, 399-428.

ROSENBERG, B. (1973b). A survey of stochastic parameter regression. *Annals of Economic and Social Measurement*, 2, 381-397.

SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHWEPPE, F. (1965). Evaluation of likelihood functions for gaussian signals. *IEE Transactions on Information Theory*, 11, 61-70.

SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (Ed.) N.K. Nawboodivi, New York: Academic Press, 201-216.

SWAMY, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.

TILLER, R. (1989). A Kalman filter approach to labour force estimation using survey data. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association* (to appear).

WATSON, M.W., and ENGLE, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23, 385-400.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 3–22
Statistics Canada

Issues and Strategies for Small Area Data

M.P. SINGH, J. GAMBINO and H.J. MANTEL[1]

ABSTRACT

This paper identifies some technical issues in the provision of small area data derived from censuses, administrative records and surveys. Although the issues are of a general nature, they are discussed in the context of programs at Statistics Canada. For survey-based estimates, the need for developing an overall strategy is stressed and salient features of survey design that have an impact on small area data are highlighted in the context of redesigning a household survey. A brief review of estimation methods with their strengths and weaknesses is also presented.

KEY WORDS: Sample design strategy; Design estimates; Model estimates.

1. INTRODUCTION

For decades, administrative records and censuses were the main sources of data used for policy and planning for both large and small areas. These are still the richest source of statistical data at small area levels in most countries. During the forties and fifties, however, as the reliance on sample surveys increased, survey based estimates complemented the traditional sources because they provide more timely and cost efficient statistical data in a variety of subject matter fields. Although designed to provide reliable estimates primarily at larger area levels such as national and provincial, increasingly such surveys are being used to meet the growing demands for more timely estimates for various types and sizes of domains. No technical problem arises as long as these domains are large enough (*e.g.*, age-sex groups, larger cities and sub-provincial regions) to yield estimates of acceptable reliability. If data are needed for small domains, however, particularly if such domains cut across design strata, special estimation problems arise and several methods have recently been proposed to deal with such problems.

The main message of this paper is to emphasize the need to look at the problem of small area data in its entirety. Small area needs should be recognized at the early stages of planning for large scale surveys. The sampling design should include special features that enable production of reliable small area data using design or model estimators. The handling of this growing challenge to statistical agencies at the estimation stage should be viewed as a last resort.

In section 2, we discuss data needs and the three main sources of socio-economic data in the Canadian context, namely, the census, administrative records and surveys. Section 3 identifies some technical issues regarding the three sources of data and highlights the problems of quality measures and their interpretation. Then a need for developing an overall strategy that includes the planning, designing and estimation stages in the survey process is highlighted in section 4. Two aspects of the design, namely, clustering in a multi-stage sample design and sample allocation are discussed. In section 5, we present some sample design options being incorporated during the current redesign of the Canadian Labour Force Survey, the largest monthly household survey conducted by Statistics Canada, with a view to enhancing the survey capacity to provide better quality small area data. The purpose of section 6 is to review the many different approaches to estimation for small areas. We also suggest some new estimators and provide comments on the strengths and weaknesses of various domain estimators. A cautious approach towards the use of model estimators is stressed.

2. INFORMATION NEEDS AND DATA SOURCES

As the country's national statistical agency, Statistics Canada plays an integral role in the functioning of Canadian society. While guaranteeing the confidentiality of individual respondents' data, the agency's information describes the economic and social conditions of the country and its people. Its economic, demographic, social and institutional statistics programs produce reliable data on many aspects of life at the national, provincial, and sub-provincial levels for use by federal and provincial governments, private institutions, academics and the media. With increases in the planning, administration and monitoring of social and fiscal programs at local levels, there has been increasing demand for more and better-quality data at these levels. Three major sources of social, socio-economic and demographic data with emphasis on small area statistics are briefly discussed below.

[1] M.P. Singh, J. Gambino and H.J. Mantel, Statistics Canada, 16th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Census of Population: The quinquennial census of population provides benchmark data and serves as the richest source of information, available every five years, for small areas and for various characteristics/domains/target groups of policy interest such as ethnic minorities, disabled persons, youth and aboriginal peoples.

Administrative Records: Administrative records are an increasingly important source of statistical data. These are extensively used in the demographic field by statistical agencies to produce local area estimates (Schmidt 1952, Verma and Basavarajappa 1987). In certain areas, such as vital statistics, administrative records are the only source of information for production of statistics at various levels of aggregation. In others, the relative merits of administrative records compared to censuses or surveys as data sources in terms of timeliness and quality of data determine the manner and the extent to which these data sources are used. In addition to direct tabulations, administrative records are used in a number of programs as a source of supplementary information for use in improving the quality of survey-based estimates. They are also being used in the construction of sampling frames for conducting surveys. Examples at Statistics Canada include the Business Register and the Address Register of residential dwellings.

Like the census of population, administrative records are very rich in geographical detail, making them a useful source of information for small area statistics. They are available more frequently and, due to recent technological advances, they are becoming a more cost-effective data source. However, administrative data are based on definitions made for programmatic rather than statistical purposes and their content is limited. Details of a Statistics Canada program for integration and development of an administrative records system to produce statistical outputs are given by Brackstone (1987a, 1987b). Experiences in the use of administrative records in other countries are included in the conference proceedings edited by Coombs and Singh (1987).

Household Surveys Program: Household surveys have long been an important source of economic and social statistics at Statistics Canada. Surveys under this program may be placed in three groups, namely, (i) the Labour Force Survey, (ii) Special Surveys and Supplementary Survey Programs and (iii) Longitudinal/Cyclical Surveys. These surveys are briefly introduced below indicating the scope for small area statistics in general.

Starting as a quarterly survey in 1945, the Canadian Labour Force Survey (LFS) became a monthly survey in 1952. The information provided by the survey has expanded considerably over the years and currently it provides a rich and detailed picture of the Canadian labour market. In addition to providing national and provincial estimates the survey regularly releases estimates for subprovincial areas. Regular estimates of standard labour market indicators are also in great demand for small areas such as Federal Electoral Districts, Census Divisions and Canada Employment Centres. These estimates are used by both federal and provincial governments in monitoring programs and allocating funds and other resources among various political and administrative jurisdictions.

Because of cost considerations, the LFS is heavily used as a vehicle for conducting *ad hoc* and periodic surveys at the national and provincial levels in the form of supplementary or special surveys. In the case of supplements, the LFS respondents themselves are asked additional questions, whereas for special surveys a different set of households is selected using the LFS frame. Both special and supplementary surveys are usually sponsored by other government departments and are conducted on a cost-recovery basis. For these surveys, the demands for small area statistics differ greatly from survey to survey, and generally the demands seem to be less pressing than those from the LFS itself.

Statistics Canada conducts a General Social Survey (GSS) annually to serve, in a modest way, the growing data needs on topics of current social policy interest. The GSS program (Norris and Paton 1991) consists of five survey cycles, each covering a different core topic, repeated every five years. Because of the limited size of sample (10,000 households nationally) the focus of the GSS is on estimates at the national level and on analytical statistics.

Longitudinal/panel surveys are new in the Canadian context. Statistics Canada has started two longitudinal surveys that will enrich the household survey program greatly, namely, the Survey on Labour and Income Dynamics and the National Population Health Survey. Both are large scale panel surveys and they are already creating expectations for data at sub-provincial and local area levels.

3. ISSUES IN DOMAIN ESTIMATION

There are numerous policy and technical issues that need to be addressed in the provision of small area statistics. The seriousness of these issues may vary from agency to agency and from one application to the next within the same agency depending on data quality and release policies. These issues are relevant for national and provincial estimates, but they assume higher significance in the context of small area statistics. As Brackstone (1987a) notes "on the issue of small area data evaluation, it is worth noting that error in small area estimates may be more apparent to users than error in national aggregates... at a local area level, there will be critics quick to point out deficiencies... it is true that for small areas, where estimation is more difficult, scrutiny of estimates is also more intensive". Several research and developmental studies on small area estimation are described in two volumes, one edited by Platek *et al.* (1987), and the other by Platek and Singh (1986). For a

recent overview of small area estimation techniques currently being used in United States federal statistical programs see U.S. Statistical Policy Office (1993).

Use of Administrative Records: Federal and provincial government policies are the prime factors that influence the supply as well as the demand for small area data in most situations. On the supply side, government program driven administrative records contain a wealth of statistical information that can be used to produce local area data. Examples of files being used in the Canadian context are: Family Allowance, Unemployment Insurance, Income Tax, Health, Education, Old Age Security. Income-related statistics are produced at the local area level on a regular basis. Any change in government policy and associated programs can have immediate impact, for better or worse, on the coverage, availability, timeliness or quality of statistics derived from the corresponding administrative records. On the demand side, as noted earlier, governments need local area data for planning, implementing and monitoring their policies.

Conceptual issues: Quite frequently, conceptual and definitional issues in a data series are confounded with sampling and estimation problems. For example, consider the Unemployment Insurance (UI) system in Canada. UI regulations stipulate different qualification and requalification periods depending on the unemployment rate in a given region such that regions with higher unemployment rates require shorter qualifying periods of continuous employment. The estimates of regional unemployment rates derived from the LFS are used in determining the eligibility for an individual to receive benefits. These local area estimates are thus continually under close scrutiny by the public and the media. Such scrutiny however refers more often to **conceptual** issues rather than estimation issues per se; aspects such as the treatment in the survey questionnaire of discouraged workers, lay-offs and job search methods are questioned.

Use of Models and Related Quality Measures: Domain estimates are produced for virtually all large scale surveys, and as long as design estimators, *i.e.*, approximately design-unbiased estimators are of acceptable quality, no problem arises. We consider two classes of design estimators. Following Schaible (1992), **direct** estimators refer to estimators which use values of the study variable only for the time period of interest and only from units in the domain (*e.g.*, the regression estimator with slope estimated using only data from the domain). Such estimators may, and often do, use information on one or more auxiliary variables from other domains or other time periods, and are design unbiased or approximately so. The second class of design estimators, **modified** direct estimators, may use information from other domains on both the auxiliary and the study variable but still retain the property of design unbiasedness or approximate unbiasedness (*e.g.*, the regression estimator with slope estimated using the whole sample). There is a growing literature on **indirect** (or **model**) **estimators**, that is, estimators which use information on both the study and auxiliary variables from outside the domain and/or the time period of interest without any reference to their design unbiasedness properties.

Most producers and users of survey data are accustomed to design estimators and the corresponding design-based inferences. They interpret the data in the context of repeated samples selected using a given probability sampling design, and use estimated design-based cvs (coefficients of variation-square root of design variance divided by the design estimate) as the measures of data quality. For situations where either domains are too small or the sampling design did not foresee production of small area estimates, the design estimates may lead to large design cvs and model estimates may be the only choice if the survey-based estimates have to be provided for individual domains. A major challenge for statisticians is how to estimate, compare and explain to the users the relative precision of estimates from a survey that produces a large number of estimates at the national, subnational and large and small domain levels, most using design estimators but a few using model estimators. The model-based cvs (square root of design variance of model estimate divided by the model estimate) may convey a completely different message and may be several times lower than the corresponding design-based cvs for the same small area and in many cases, lower than the design-based cvs for much larger areas.

For model estimators, it is usually straightforward to derive expressions for the corresponding mean square errors (*i.e.*, design variance + square of the design bias). Estimation of these expressions, with an adequate degree of reliability, is a different matter. If we follow the argument that the data (*e.g.*, sample size) for such domains are inadequate for producing design estimates, it is unlikely that they would be adequate for producing design estimates of the corresponding variances and biases. As the estimation of bias is relatively more difficult, some authors seek design consistent model estimators, implying perhaps that bias can be ignored. However, if the sample size within the domain is sufficiently large to make the model estimator consistent, then the design estimator itself should give reliable estimates for the domain. For model estimators, suggestions have been made to use estimates of average mean square error computed over all domains. As the need for estimates for different domains usually arises because these domains are thought to be different from each other, a challenging task is to explain why estimates from all such domains are given the same degree of reliability. Another possibility is to construct indirect model-based estimates of the variance and bias of the model estimators for **individual** domains. Finding suitable methods of estimating mean square error for individual domains should be a research priority. Another serious concern for survey practitioners is how to guard against model failures. This

suggests a need for research into model validation for complex survey situations. Further, for model estimators that use data on study variables for periods other than the time period of interest, estimates of change over different time periods would be of questionable quality; see Schaible (1992). Also, model estimators that borrow strength from other domains in the larger area will suffer a similar drawback when comparing differences in the two domains within the large area.

Issue of Privacy: In order to construct rich data bases for providing small area statistics, it is sometimes necessary to combine census, survey and/or administrative records. This necessitates linkage of records obtained from different sources. However, given the public's concern about privacy, record linkages should be carried out only after careful examination of all their implications. Under the Statistics Act, Statistics Canada may have access to administrative records of other departments for statistical purposes. But even for statistical purposes, as Fellegi (1987) notes, "we should have rigorous and auditable review procedures to ensure that we only carry out record linkage where the resulting privacy invasion is clearly outweighed by the public good from the new statistical information".

4. NEED FOR AN OVERALL STRATEGY

Even though large scale surveys are designed primarily for national and provincial estimates, it is rare that the estimates from such surveys relate only to the national/provincial populations as a whole. That is, invariably, such surveys are used to produce estimates for various cross-classified domains and in some cases for areal domains (*e.g.*, subprovincial) as well. In many cases, no special attention is paid to achieving a desired level of precision at the domain level either at the design or the estimation stage as long as the reliability is (believed to be) within reasonable limits. Problems arise when the cross-classified domain refers to a rare subpopulation or when the areal domain refers to a small area in which case either no estimates are possible/available or the estimates are of questionable quality. In a number of cases, this may happen simply because not enough attention was paid to these needs at the start of the survey planning process. If small area data needs are to be served using survey data then there is a need to develop an overall strategy that involves careful attention to meeting these needs at the planning, sample design and estimation stages of the survey process. For discussion of the design and estimation aspects, we will classify domains into the following two types:

Planned domains: In sampling terms these are individual strata or groups of strata for which desired samples have been planned. In the Canadian context these are typically subprovincial regions, such as Economic Regions, Unemployment Insurance Regions, and Health Planning Regions. In other cases, such domains could be larger counties, districts or similar subprovincial regions.

Unplanned domains: These are areas that were not identified at the time of design and thus may cut across design strata. Such domains can be of any size and they may create special estimation problems.

Planning: As noted earlier, the data demands from continuing periodic surveys such as the LFS are relatively much higher than from *ad hoc* surveys. In the case of periodic surveys that are redesigned every five or ten years, a suitable strategy can be developed during survey redesigns, since, in such cases, statistical agencies are usually in a much better position to project future small area data needs based on past demands. For *ad hoc* surveys, designers should include the establishment of such needs as an integral part of objective setting for the survey. Thus, in both cases, survey designers should establish the desired degree of precision, not only for national and provincial level estimates, but also for the domains of interest.

The first step of a strategy, in terms of the provision of small area data, will depend on the extent to which domains are identified in advance so that they can be treated as planned domains at the time of the design (or redesign) of the survey. If budgetary considerations do not permit reliable estimates for certain very small domains, then the option of either collapsing domains, pooling estimates over different surveys or not providing the estimates at all should be given serious consideration by survey designers in discussions with the survey sponsors. Some domains cannot be determined in advance. These unplanned domains should be handled through special estimation methods.

Sample design: In practice, it is rare that a design is optimal either for the national or provincial levels or for a single subject matter of interest. Usually varying degrees of compromise are introduced at different stages of sampling and the data collection process to satisfy theoretical and operational constraints. Depending on the data needs, estimates for domains should also form an integral part of this compromise. We will discuss two ways of taking small area data needs into account at the design stage, namely, sample allocation and the degree of clustering of the sample.

Allocation Strategy: In general, an optimum allocation strategy for national level estimates allocates samples to provinces approximately in proportion to their population. The reliability of estimates for smaller provinces in such cases suffers. Therefore a compromise allocation is usually preferred. There are different ways in which this compromise can be achieved depending on the emphasis placed on subnational estimates. Small reductions in sample sizes for larger provinces usually have little effect or the reliability of data for such provinces (or the national level data) but the corresponding sample increase in smaller provinces has significant impact on the reliability of their data.

The same principle holds for planned domains within the provinces. This is because optimum allocations in most situations are flat and the designers can exploit this feature by reallocating sample from the larger areas to planned domains that are smaller in size.

Clustering: Large scale household surveys usually involve stratified multistage designs with relatively large primary sampling units in order to make the design cost-efficient for national and provincial statistics. Such designs are thus highly clustered and, therefore, detrimental to the production of statistics for unplanned areal domains in the sense that, due to chance, some domains may be sample-rich while others may have no sample at all. Given the importance of domain estimates, attempts should be made to minimize the clustering in the sample. The following factors are important in this context: choice of frame, choice of sampling units and their sizes, number and size of strata and stages of sampling. The goal should be to make the design effects as low as possible given the operational constraints.

Estimation: No matter how much attention is paid to domain estimates at the early stages of planning and designing a particular survey, there will always be some smaller domains for which special estimation methods will be required for producing adequate estimates. Recently, synthetic estimators, which borrow strength from domains that resemble the domain of interest, have attracted a good deal of attention. However, since synthetic estimators are very sensitive to the assumption that domains resemble each other, even a small departure from the assumption can make the design bias high and put their use in question. Probability samplers, conscious of design bias, have suggested combinations of direct and synthetic estimators, with a view to addressing the design bias problem while trying to retain the strengths of the synthetic estimator. Empirical Bayes and similar techniques have been used to assign a weight to each component in the combined estimators. A brief review of these developments is given in section 6 on estimation.

5. SAMPLE DESIGN CONSIDERATIONS

5.1 Introduction

The small area problem is usually thought of as one to be dealt with via estimation. However, as was noted in the previous section, there are opportunities to be exploited at the survey design stage. This section uses the Canadian Labour Force Survey (LFS) to illustrate this.

The current LFS design: The Canadian Labour Force Survey is a monthly survey of 59,000 households which are selected in several stages using various methods. The ultimate sampling unit, the household, remains in the sample for six months once it is selected and is then replaced. Higher stage units (primary sampling units (PSU), clusters) also rotate periodically. Each of Canada's ten provinces is divided into economic regions (ER) which the LFS further divides into self-representing areas (medium and large cities) and non-self-representing areas (the rest of the ER). Stratification and sample selection take place within these areas, and the number of stages of sampling as well as the units of sampling differ between these two types of area. For example, in areas outside cities, there are three stages of sampling, whereas there are only two in the cities. For a detailed description of the current LFS design, refer to Singh *et al.* (1990).

5.2 Sampling Stages and Sampling Units

Area frames are usually associated with clustered sampling, *i.e.*, the first-stage units of selection are typically land areas containing a number of second-stage units. If a list of the second-stage units becomes available, then sampling directly from the list becomes possible, leading to a less clustered sample. This will result not only in improved estimates (due to lower design effects) but also in better small area estimates for unplanned domains. The latter holds since, by spreading the sample more evenly, it is more likely that an unplanned areal domain will contain some selected units. In contrast, in a clustered design we are often faced with a situation where one domain has sufficient sample because it happens to contain sampled clusters while a similar domain happens to have too few or no sampled clusters to produce good estimates.

To reduce clustering in the LFS we investigated two options: (i) the possibility of replacing the area frame (with its two stage design) in the larger cities with a list frame using the Address Register and (ii) reducing the sampling stages in rural areas and smaller urban centres. The Address Register, created to improve the coverage of the 1991 Canadian census (Swain, Drew, Lafrance and Lance 1992), consists of a list of addresses, telephone numbers and geographical information for dwellings by census enumeration area (EA). One option involved the selection of a stratified simple random sample of dwellings from the Address Register frame. This sample could then be supplemented with a sample selected from a growth frame which comprises a set of dwellings that are not in the post-censal address register. Handling of growth became the major stumbling block in pursuing option (i) as no cost-efficient method could be devised and tested in time for the current redesign. However, an updating strategy for the post-censal Address Register is still being investigated for future censuses and surveys.

With regard to option (ii), in keeping with the idea that less clustering is better for small area estimates, changes in the units and reduction in the stages of sampling were investigated for the areas outside the cities. Due to the changes that have taken place in data collection techniques,

namely, from face-to-face interviewing to telephone and computer assisted interviewing, the cost-variance analyses from the past are no longer relevant. More than 80 percent of LFS interviews are now conducted by telephone. With the increase in telephone interviewing and the resulting decrease in travel, it became feasible in almost all cases to eliminate the current PSU stage and to sample EAs directly.

5.3 Stratification

One approach to stratification, similar in spirit to the above discussion on PSU size, is to replace large strata by many small ones. The hope is that a redefined domain or an unplanned domain will contain mostly complete strata. This will make the sample size in the domain more stable.

There may be several overlapping areas for which estimates are required. For example, each Canadian province is partitioned into both Economic Regions (ER) and Unemployment Insurance regions (UIR). One way to deal with this situation is to treat all the areas created by the intersections of the partitions as strata. In the Canadian case, for example, the 71 ERs and 61 UIRs yield 133 intersections, a manageable number. In some cases, however, the number of intersections may be too large to handle effectively. In addition, some of the intersections may have very small populations, making them unusable as strata.

By combining decreased clustering with smaller strata, we hope to have a design which is better able to meet small area needs. For example, the design should provide more flexibility in satisfying both ER and UIR requirements efficiently and in dealing with future changes in the definition of regions.

5.4 Allocation

If the definitions of small areas are known in advance, we may be able to treat them as planned domains and take them into account when designing the survey. The survey designer may endeavour to allocate sufficient sample in each small area to make the production of reliable estimates feasible. For large surveys such as the Canadian Labour Force Survey, this approach can, at least in theory, make the production of a great many small area estimates feasible. With a monthly sample of 59,000 households, and assuming that, say, 100 households per month are needed to produce reliable quarterly estimates, the country can be divided into about 600 non-overlapping areas, each guaranteed to have sufficient sample. Unions of such areas will also have enough sample to produce reliable monthly estimates.

Various sample allocation strategies are possible. In a top-down approach, once a provincial sample size is determined, the sample is allocated among the sub-provincial regions. However, it may turn out that it is not possible to satisfy the requirements for the reliability of sub-provincial estimates for the given provincial sample size. In a bottom-up strategy, the sample would be allocated to sub-provincial regions first in such a way that reliability objectives for each region are satisfied. As a result, we would expect comparable sample sizes in each sub-provincial region. This approach may result in a provincial sample size that is bigger than the one specified in the top-down approach. Regardless of which of the two strategies is used, adjustments to the initial allocations will usually be required. The resulting allocation will likely resemble a compromise between proportional allocation and equal allocation. In practice, the survey designer must perform a complex juggling act among provincial reliability requirements, sub-provincial requirements for one or more sets of regions, total survey costs and in-the-field details.

The approach taken in the current LFS redesign may be useful in other surveys as well. The sample was allocated in two steps: first, a core sample of 42,000 households was allocated to produce good estimates at the national and provincial levels; then the remaining sample was allocated to produce the best possible sub-provincial estimates. The resulting compromise allocation will produce reliable estimates for almost all planned domains. The compromise resulted in only minor losses at the provincial level and substantial gains at the subprovincial level. For example, the expected CVs for 'unemployed' for Ontario and Quebec are 3.2 and 3.0 per cent, respectively, instead of 2.8 and 2.6. The corresponding figures for Canada are 1.51 and 1.36. Optimizing for the provincial level yields CVs as high as 17.7 per cent for UI regions. With the compromise allocation, the worst case is 9.4 per cent.

Sample redistribution: There is usually some scope for moving sample from one area to another. For example, reducing the sample size by 1,000 households in a large province and making a corresponding increase in a small province will cause a marginal deterioration in the quality of provincial estimates in the former but will improve the estimates in the latter significantly. Similar movements of sample can be attempted within province.

5.5 Other Considerations

Change in definitions of small areas: Survey designers are faced with the fact that the definitions of planned domains may change during the life of a design and they may then have to treat the new domains as unplanned domains. For example, it is quite possible that the definitions of Unemployment Insurance Regions will change two or three years after the new LFS design is introduced in 1995. To deal with this at the design stage, the best that the survey designer can do is to choose as building blocks areas which are standard (*e.g.*, census-defined areas whose definitions are fairly stable) and hope that the redefined regions are unions of these standard areas. This is the approach that was taken in the current LFS redesign.

An alternative is to adopt an update strategy. This entails a reselection of units, doing it in such a way that the overlap between the originally selected units and the newly selected ones is maximized. By taking this approach, the number of new units that have to be listed is minimized. This also minimizes other field disruptions such as the need to hire new interviewers.

6. ESTIMATION

The purpose of this section is to review some of the different approaches to estimation of totals for small areas. No attempt is made to provide an exhaustive review; the discussion indicates the trend of developments in small area estimation research. For a detailed review, see the recent paper by Ghosh and Rao (1993). To facilitate this review we will classify small area estimation methods into two types. This is just one of many possible classification schemes. The first class of estimators we call design estimators, *i.e.*, (approximately) design unbiased estimators, which includes direct and modified direct estimators. As noted earlier, design estimators are often unsatisfactory, having a large variance due to small sample sizes (or even no sample at all) in the small areas. The second class we call indirect (or model) estimators, and it includes synthetic and combined estimators. Some of these estimators are compared empirically in an earlier version of this paper by Singh, Gambino and Mantel (1992).

6.1 Design Estimators

Direct Estimators: Direct small area estimators are based on survey data from only the small area, perhaps making use of some auxiliary data from census or administrative sources in addition to the survey data. The simplest direct estimator of a total is the expansion estimator,

$$\hat{Y}_{e,a} = \sum_{i \epsilon s_a} w_i y_i, \qquad (6.1)$$

where s_a is the part of the sample in small area a and w_i is the survey weight for unit i. This estimator is unbiased; however, it may have high variability due to the random sample size in area a.

If the population size N_a is known then a post-stratified estimator,

$$\hat{Y}_{pst,a} = N_a \sum_{i \epsilon s_a} w_i y_i \Big/ \sum_{i \epsilon s_a} w_i = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \bar{y}_{e,a}, \qquad (6.2)$$

may be used. This estimator is more stable than the expansion estimator; however, there may be some ratio estimation bias in complex surveys.

If the sampling scheme is stratified and the $N_{h,a}$ are known, where $N_{h,a}$ is the population size in stratum h and small area a, an alternative post-stratified estimator is $\hat{Y}_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \epsilon s_{h,a}} w_i y_i / \sum_{i \epsilon s_{h,a}} w_i) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \bar{y}_{h,a}$. The strata may also be post-strata instead of design strata.

Ratio estimation is similar to post-stratified estimation, the difference being that another auxiliary variable is used in place of the population counts N_a and $N_{h,a}$. For example, if x is a covariate for which the small area totals, X_a, or the stratum small area totals, $X_{h,a}$, are known then we may define the ratio estimators

$$\hat{Y}_{r,a} = X_a \hat{R}_a \quad \text{and} \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a}, \qquad (6.3)$$

where $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$ is an estimate of the ratio Y_a / X_a and $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$.

A regression estimator attempts to account for differences between small area subpopulation and subsample values of the covariates via an estimated regression relationship between the variate of interest, y, and the covariates, x. An advantage of regression type estimation is that it is easily extended to vector covariates. The estimator is given by

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a), \qquad (6.4)$$

where $\hat{Y}_a$ may be an expansion or post-stratified estimator, $\hat{X}_a$ must be calculated in the same way as $\hat{Y}_a$, and $\hat{\beta}_a = \sum_{i \epsilon s_a} v_i^{-1} w_i y_i x_i' \{ \sum_{i \epsilon s_a} v_i^{-1} w_i x_i x_i' \}^{-1}$ where v_i are given weights for the regression. Note that $\hat{\beta}_a = \hat{R}_a$ when x is scalar and $v_i = x_i$. When $\hat{Y}_a$ and $\hat{X}_a$ are expansion estimators this estimator is also called the generalized regression estimator. Approximate design unbiasedness of this estimator follows from that of $\hat{Y}_a$ and $\hat{X}_a$.

As with the ratio type estimators, regression type estimation may also be applied within design strata or post-strata.

Modified Direct Estimators: Modified direct estimators may use survey data from outside the domain; however, they remain approximately design unbiased. By a modified direct estimator we mean a direct estimator with a synthetic adjustment for model bias; since the adjustment would have approximately zero expectation with respect to the design, the modified estimator is approximately design unbiased if the direct estimator is. An example is obtained by replacing $\hat{\beta}_a$ in (6.4) by a synthetic estimator $\hat{\beta} = \sum_{i \epsilon s} v_i^{-1} w_i y_i x_i' \{ \sum_{i \epsilon s} v_i^{-1} w_i x_i x_i' \}^{-1}$; we will denote this estimator by $\hat{Y}_{sreg,a}$. $\hat{\beta}$ would generally be more stable than $\hat{\beta}_a$; the choice between them would depend on the size of the variance of $\hat{\beta}_a$ relative to the variation in the β_as over areas a. A compromise is to take a weighted average $\lambda_a \hat{\beta}_a + (1 - \lambda_a) \hat{\beta}$ where λ_a is suitably chosen;

options for the choice of λ_a are discussed under combined estimators in Section 6.2. A second example is obtained by replacing $\hat{\beta}_a$ in (6.4) by $\hat{R} = \hat{Y}_e/\hat{X}_e$; note that $\hat{R}$ is a special case of $\hat{\beta}$ where x is scalar and $v_i = x_i$.

6.2 Indirect Estimators

Synthetic Estimators: Synthetic estimation methods are based on an assumption that the small area is similar in some sense to another area, often a larger area which contains it. Estimates for the other area would generally be more reliable than those for the small area. The resulting synthetic estimator would then have small variance, though it may be badly biased if the underlying assumption is violated.

One of the simplest synthetic estimators arises from the assumption that the small area mean is equal to the overall mean. This leads to the mean synthetic estimator

$$\hat{Y}_{syn,m,a} = N_a \sum_{i\epsilon s} w_i y_i \Big/ \sum_{i\epsilon s} w_i = N_a \bar{y}. \qquad (6.5)$$

A more common synthetic estimator is based on stratification or post-stratification,

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \sum_{i\epsilon s_h} w_i y_i \Big/ \sum_{i\epsilon s_h} w_i = \sum_h N_{h,a} \bar{y}_h.$$

As with direct estimators, ratio synthetic estimation may be based on other auxiliary data besides the population counts N_a or $N_{h,a}$. For example, the common ratio synthetic estimators based on a covariate x are defined as

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e/\hat{X}_e \quad \text{and} \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \hat{Y}_{h,e}/\hat{X}_{h,e}, \qquad (6.6)$$

where $\hat{Y}_e = \sum_{i\epsilon s} w_i y_i$ is the expansion estimator of the population total for y and $\hat{Y}_{h,e} = \sum_{i\epsilon s_h} w_i y_i$. $\hat{X}_e$ and $\hat{X}_{h,e}$ are similarly defined. These estimators have been studied by Gonzalez (1973), Gonzalez and Waksberg (1973) and Ghangurde and Singh (1977, 1978), among others.

Singh and Tessier (1976) suggested an alternative ratio synthetic estimator, using X instead of $\hat{X}_e$, defined as

$$\tilde{Y}_{syn,r,a} = X_a \hat{Y}_e/X. \qquad (6.7)$$

Both $\hat{Y}_{syn,r,a}$ and $\tilde{Y}_{syn,r,a}$ have the same synthetic bias and the ratio bias in $\hat{Y}_{syn,r,a}$ will be negligible for large samples. The choice between these two estimators depends on ρ, the correlation of $\hat{Y}_e$ and $\hat{X}_e$. It can be shown that for large samples $V(\hat{Y}_{syn,r,a}) \leq V(\tilde{Y}_{syn,r,a})$ if $\rho \geq 0.5 c_x/c_y$, where c_x and c_y are the coefficients of variation of $\hat{X}_e$ and $\hat{Y}_e$, respectively. In most cases, when ρ is high or the population is skewed, $\hat{Y}_{syn,r,a}$ would be preferred; however, when c_x is high and the correlation is only moderate, $\tilde{Y}_{syn,r,a}$ may be the better choice.

In some situations information on a second auxiliary variable (z) in addition to x may be available. Then a bivariate ratio synthetic estimator may be constructed:

$$\hat{Y}^{(2)}_{syn,r,a} = \gamma_a X_a \hat{Y}_e/\hat{X}_e + (1 - \gamma_a) Z_a \hat{Y}_e/\hat{Z}_e, \qquad (6.8)$$

where γ_a is suitably chosen. Extensions to a multivariate ratio synthetic estimator may be considered following Olkin (1958).

Regression synthetic estimation is similar to ratio synthetic,

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a,$$

$$\hat{\beta} = \sum_{i\epsilon s} v_i^{-1} w_i y_i x_i' \left\{ \sum_{i\epsilon s} v_i^{-1} w_i x_i x_i' \right\}^{-1}. \qquad (6.9)$$

Again, regression synthetic estimation may also be applied within design strata or post-strata. Royall (1979) suggested a slight variation, $\hat{Y}_{syn,Roy,a} = \sum_{i\epsilon s_a} y_i + \hat{\beta}(X_a - \sum_{i\epsilon s_a} x_i)$, where the sum of y-values for only units not included in the sample is estimated synthetically.

Remark: The examples of modified direct estimators presented in Section 6.1 can also be considered to be ratio or regression synthetic estimators with a design-based adjustment to correct for bias. For example, we may write $\hat{Y}_{sreg,a} = \hat{Y}_{syn,reg,a} + (\hat{Y}_a - \hat{\beta}\hat{X}_a)$ where $\hat{Y}_a - \hat{\beta}\hat{X}_a$ is an estimate of the bias of $\hat{Y}_{syn,reg,a}$. Similarly, $\hat{Y}_{sreg,a}$ can also be written as the Royall estimator, $\hat{Y}_{syn,Roy,a}$, with a design-based adjustment for bias.

Purcell and Kish (1980) discuss another type of synthetic estimation which they call SPREE (structure preserving estimation) for small area estimation of frequency data. Detailed historical counts, perhaps from a census, are combined with less detailed current survey estimates to produce detailed estimates of current counts. The assumption here is that certain relationships among the detailed counts are stable over time.

Combined Estimators: By a combined estimator we mean a weighted average of a design estimator and a synthetic estimator,

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{des,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \qquad (6.10)$$

where λ_a is suitably chosen. The aim here is to balance the potential bias of the synthetic estimator against the instability of the design estimator. There are three broad approaches which may be used to define the weights λ_a in (6.10); they may be fixed in advance, sample size dependent, or data dependent.

The first and simplest approach to weighting is to fix the weights in advance, for example, to take a simple average. However, this does not make any allowance for

the actual observed reliability of the design estimator. For some realized samples the design estimator for small area a is more reliable than for other realized samples. The weight given to the design estimator should reflect this.

The second general approach to weighting of the design and synthetic parts is called sample size dependent, in which the weights are functions of the ratio $\hat{N}_{e,a}/N_a$. Another possibility, not considered here, is to base the weights on the realized sample values of a covariate x; for example, the weight could be a function of $\hat{X}_{des,a}/X_a$ or of $S^2_{x,a}/\sigma^2_{x,a}$ where $S^2_{x,a}$ is the realized variance of $\hat{X}_{des,a}$, conditional on $\hat{N}_{e,a}$ or some other relevant aspect of the realized sample, and $\sigma^2_{x,a}$ is the unconditional variance of $\hat{X}_{des,a}$.

Some specific estimators in this class have been proposed earlier. Drew, Singh, and Choudhry (1982) proposed the sample size dependent estimator

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1 - \lambda_a)\hat{Y}_{syn,r,a}, \qquad (6.11a)$$

where

$$\lambda_a = \begin{cases} 1 & \text{if} \quad \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & \text{otherwise} \end{cases} \qquad (6.11b)$$

and δ is subjectively chosen to control the contribution of the synthetic component. Särndal (1984) suggested

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1 - \lambda_a)\hat{Y}_{syn,reg,a}, \qquad (6.12)$$

where $\lambda_a = \hat{N}_{e,a}/N_a$. Rao (1986) suggested a modification to this in which λ_a would be taken to be 1 whenever $\hat{N}_{e,a} \geq N_a$. Särndal and Hidiroglou (1989) refined Rao's suggestion by taking $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ when $\hat{N}_{e,a} < N_a$, where h is chosen judgementally to control the contribution of the synthetic component.

It is the bias of the synthetic component that is of concern when using these sample size dependent estimators in practice. The weight associated with the synthetic component should be such that the bias is kept within reasonable limits. For example, the sample size dependent estimator of Drew, Singh and Choudhry (1982), with generalized regression estimation replacing the ratio estimation and $\delta = 2/3$, is currently used in the Canadian Labour Force Survey to produce domain estimates. For a majority of domains the weight attached to the synthetic component is zero as the direct estimator itself provides the required degree of reliability. For other domains the weight attached to the synthetic component is about 10% on average and never exceeds 20%. Depending on the risk of bias that one is willing to take, δ may lie in the range [2/3,3/2] for most practical situations.

The third approach to weighting we call data dependent. The optimal weights for combining two estimators generally depend on the mean squared errors of the estimators and their covariance. These quantities would generally be unknown but may be estimated from the data. For our combined estimators this would usually require some modelling of the bias of the synthetic part. An early and well known example of this approach is due to Fay and Herriot (1979). They model the biases of the synthetic estimators for the small areas as independent random effects with an unknown but fixed variance. To be more specific, if $\hat{Y}_{des,a}$ is the design estimator then they consider the model $Y_a = X_a\beta + \alpha_a$ and $\hat{Y}_{des,a} = Y_a + \epsilon_a$ where $\alpha_a \sim (0,\sigma^2)$, $\epsilon_a \sim (0,\nu_a^2)$, and α_a and ϵ_a are independent and uncorrelated over a, σ^2 is unknown and ν_a^2 are assumed known (in practice they would need to be estimated). For a given value of σ^2 the optimal weights for combining $\hat{Y}_{des,a}$ and $X_a\hat{\beta}$ can be calculated. An estimate of σ^2 is obtained by the method of fitting constants and substituted into the optimal weights. Some protection against model mis-specification is obtained by truncating the resulting estimate if it deviates from the direct estimate by more than a specified multiple of ν_a. Schaible (1979) and Battese and Fuller (1981) also consider empirically estimated optimal weights λ_a in (6.12) based on similar random effects models for the small area totals.

Prasad and Rao (1990) provide an estimator of the mean square error of the Fay-Herriot estimator which makes allowance for the estimation of the variance components. Kott (1989) proposes a design consistent estimator of the mean square error, but finds it to be very unstable.

Another alternative is to use historical data to calculate the weights; this has the advantage that the weights may be more stable than if they are estimated from current survey data; however, there is an underlying assumption that the optimal weights are stable over time.

Remark: In sample size dependent estimation the weights are allowed to depend on the observed size of the subsample s_a, but not on the values of the variate of interest. This non-dependence of the weights on the variate of interest has advantages and disadvantages. An advantage is that the same weights would be used for estimation of totals for all variates of interest; they need to be calculated only once. More importantly, the estimate of the sum of two variables is the sum of the estimates of the two variables. A disadvantage is that the weights do not directly take account of either the reliability of the design estimator for the variate of interest or the likely magnitude of the bias of the synthetic estimator.

Combining data over time: For repeated surveys pooling of data over survey occasions to increase the reliability of estimates is a common practice. Depending on the rotation pattern used for such surveys, significant gains in reliability can be achieved. This pooling or averaging over time is thus of particular interest in the context of domain estimation where reliability is usually low. For domain

estimation in the Canadian Labour Force Survey it is normal practice to use a sample size dependent estimator based on three month average estimates of employed and unemployed. Due to the six month rotation scheme used, as noted earlier, averaging over three months increases the sample size by one third. If samples completely overlap between periods then averaging does not result in any gain in efficiency. For other rotation patterns the sample size for domain estimates could be more than doubled through this process. There is, however, a conceptual problem with pooled estimates, in that such estimates refer to an average of the parameter of interest (*e.g.*, unemployment) over a period of, say, three months.

In composite estimation the current design estimator is combined with the composite estimator for the previous period, updated by an estimate of change based on the common sample. This idea was used, though not in the context of small area estimation, by Jessen (1942), and Patterson (1950), among others. Binder and Hidiroglou (1988) provide a review. The weights for the combination are typically estimates of the optimal weights under the assumption that these weights are time stationary. These data dependent weights have the disadvantage that they lead to inconsistency of estimates for different characteristics and their sums.

A recent development in small area estimation techniques is the use of time series methods for periodic surveys. The relationship between parameters of interest for different time periods is modelled and this model is exploited to improve the efficiency of the estimates for the current occasion. In most cases some allowance must also be made, through modelling or otherwise, for the non-independence of samples for different survey occasions due to the sample rotation scheme. Some references for this time series approach are Choudhry and Rao (1989), Pfeffermann and Burck (1990), Singh, Mantel and Thomas (1994) and Singh and Mantel (1991). All of these are generalizations of the Fay-Herriot model which allow the regression parameters, small area effects, and survey errors to evolve over time according to various time series models. The vector of small area estimates that results from this approach can be written as a weighted average of the vector of design estimates and a vector of synthetic estimates which are based on past data and the current values of covariates; however, the matrix of weights would not generally be diagonal so that the estimator for any single small area would generally depend also on the design estimates and synthetic estimates for other small areas.

7. CONCLUSION

To produce adequate survey-based domain estimates that are timely and up to date, sample designers must face several challenging tasks. The first is to convince the sponsors/program managers that some small area data needs cannot be met as a by-product of a system designed optimally for national/sub-national estimates. Significant gains, which may vary from survey to survey, can be achieved at the domain level at a marginal reduction in reliability at higher levels. There is a need to develop an overall strategy that incorporates desired reliability for the planned domains as well as for higher levels through compromise allocations, and reduced clustering to help improve estimates for unplanned domains. It should be noted that many of the planned domains at design time may become unplanned (revised) over time in the context of continuous surveys.

The overall strategy should also include consideration of both design estimators for larger domains and model estimators for small domains. A model estimator should be preferred over a design estimator only if its mean square error (design variance + bias^2) is estimable and it is sufficiently smaller than the corresponding variance of the design estimator. We should have estimates of mean square error for each of the individual domains. An option that statistical agencies can exercise is to pool similar domains or pool estimates over different time periods for the same domain. They may even suppress estimates for some domains on account of data reliability or privacy concerns.

The second challenging task for statisticians is to explain to users the different types of measures of reliability for different sets of estimates from the same survey. It is hoped that with more research on model validation and better estimates of mean square errors, designers will get more confidence in using model estimators for small domains. In the meantime model estimators should be used with caution even if they have significantly smaller coefficients of variation.

Censuses, supplemented by data from administrative records, are likely to remain the primary source of small area socio-economic data, especially for countries having a quinquennial census of population and housing. Also, concerns about problems with conceptual issues in the context of data for administrative records are likely to continue until statistical agencies are given an opportunity to influence the development of the forms used to collect such data. Until then, this immensely rich data source cannot be fully exploited for statistical purposes and more so for domain estimation.

ACKNOWLEDGMENT

We are grateful to Jon Rao for handling this paper as an Associate Editor and to the referees for many constructive suggestions.

REFERENCES

BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BINDER, D., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Eds. P.R. Krishnaiah and C.R. Rao). New York: Elsevier Science, 187-211.

BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 3-20.

BRACKSTONE, G.J. (1987b). Statistical uses of administrative data: Issues and challenges. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 5-16.

CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings: Symposium on Analysis of Data in Time*, (Eds. A.C. Singh and P. Whitridge), Statistics Canada, 67-74.

COOMBS, J.W., and SINGH, M.P. (Eds.) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLEGI, I.P. (1987). Opening Remarks. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 1-2.

GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. To appear in *Statistical Science*.

GHANGURDE, P.D., and SINGH, M.P. (1977). Synthetic estimates in periodic household surveys. *Survey Methodology*, 3, 152-181.

GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.

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

GONZALEZ, M.E., and WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.

NORRIS, D., and PATON, D. (1991). Canada's General Social Survey: Five years of experience. *Survey Methodology*, 17, 227-240.

OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. Invited Presentations. New York: Wiley.

PLATEK, R., and SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (contributed papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48, 3-18.

RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.

ROYALL, R.M. (1979). Prediction models in small area estimation. In *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare.

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

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. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress catalogue number 79-600067, 36-53.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 95-114.

SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.

SINGH, A.C., and MANTEL, H.J. (1991). State space composite estimation for small areas. *Proceedings: Symposium 91, Spatial Issues in Statistics*, Statistics Canada, 17-25.

SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

SINGH, M.P., GAMBINO, J.G., and MANTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 37-75.

SINGH, M.P., and TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register for coverage improvement in the 1991 Canadian Census. *Survey Methodology*, 18, 127-142.

U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Prepared by the subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology.

VERMA, R.B.P., and BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 46-61.

COMMENT

W.A. FULLER[1]

The authors are to be congratulated on an excellent description of the design and estimation considerations associated with domains. The authors discuss estimation for planned domains, particularly situations in which domain membership can be identified in the frame, and estimation for unplanned domains including domains for which the domain membership cannot be determined from the frame. This is a fine contribution to the growing literature on domain estimation.

The authors give a particularly good description of the planning, data collection, and processing activities associated with surveys conducted by Statistics Canada. Included are the traditional design problems of balancing needs for domain estimation with desire for efficiency at higher levels, the importance of confidentiality in using administrative records in constructing domain estimates, and the importance of definitional compatibility in attempting to combine information from different sources.

The importance of considering domain estimation at the design stage is very well taken and is a point often ignored by authors concentrating on small area estimation. As the authors emphasize, careful design can often enable one to construct estimates for domains in a direct and design consistent manner. I am sure that those actually designing surveys have considered the importance of clustering when designing surveys that will be used for domain estimation, but it is pleasant to see an explicit discussion.

The authors describe several types of estimators for domains. Their classification emphasizes the number of alternatives available to the practitioner. It is possible to use the theoretical mean square errors to provide information on the relative merits of the estimators. As an example of such a comparison, assume a simple random sample of size n selected from a population divided into K domains. Assume that the domain sizes and the domain means of an auxiliary variable, X, are available. Consider the three regression estimators of the domain mean,

$$\hat{\mu}_{(1)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_i,$$

$$\hat{\mu}_{(2)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_.$$

and

$$\hat{\mu}_{(3)yi} = \bar{y}_{..} + (\mu_{xi} - \bar{x}_{..})b_.,$$

where

$$(\bar{x}_{..}, \bar{y}_{..}) = \sum_{i=1}^{k} N^{-1}N_i(\bar{x}_{i.}, \bar{y}_{i.}),$$

$$(\bar{x}_{i.}, \bar{y}_{i.}) = n_i^{-1} \sum_{j=1}^{n_i} (X_{ij}, Y_{ij}),$$

$$b_i = \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

$$b_. = \left[\sum_{i=1}^{k} N^{-1}N_i n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{i=1}^{k} N^{-1}N_i n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

n_i is the number of observations in domain i, N_i is the population size of domain i, μ_{xi} is the population mean of X for domain i, and $\mu_{x.}$ is the grand population mean of X. In the authors' terminology, the first estimator is a direct regression estimator, the second is a modified direct estimator, and the third is a synthetic estimator. We have

$$\text{MSE}\{\hat{\mu}_{(1)yi} | n_i\} = n_i^{-1}(1 + n_i^{-1}) V\{Y_{tj} - \beta_t X_{tj} | t = i\} + O(n_i^{-2}),$$

$$\text{MSE}\{\hat{\mu}_{(2)yi} | n_i\} = n_i^{-1}(1 + n^{-1}) V\{Y_{tj} - \beta X_{tj} | t = i\} + O(n^{-2}),$$

$$\text{MSE}\{\hat{\mu}_{(3)yi} | n_i\} = (1 + n^{-1}) \times \sum_{i=1}^{k} N^{-2}N_i^2 n_i^{-1} V\{Y_{tj} - \beta X_{tj} | t = i\} + (\mu_{xi} - \mu_{x.})^2 V\{b_.\} + [\mu_{yi} - \mu_{y.} - \beta(\mu_{xi} - \mu_{x.})]^2 + O(n^{-2}),$$

where $V\{b_.\} = E\{(b_. - \beta)^2\}$, $V\{a_t | t = i\}$ is the variance of the variable a for domain i,

$$\beta_i = [V\{X_{tj} | t = i\}]^{-1} C\{Y_{tj}, X_{tj} | t = i\}$$

[1] W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

and

$$\beta = \left[\sum_{i=1}^{k} N^{-1} N_i V\{X_{\ell j} \mid \ell = i\} \right]^{-1} \times \sum_{i=1}^{k} N^{-1} N_i C\{Y_{\ell j}, X_{\ell j} \mid \ell = i\}.$$

The estimator $\hat{\mu}_{(1)yi}$ uses only information in the sample of n_i observations. Hence, all properties of the estimator are functions of n_i and of the domain parameters. The regression bias is order n_i^{-1} and the variance is order n_i^{-1}. The estimator $\hat{\mu}_{(2)yi}$ uses the domain means, but the entire sample to estimate the regression coefficient. Hence, the basic variance remains order n_i^{-1} and will be larger than the basic variance of $\hat{\mu}_{(1)yi}$ in those situations where $\beta_i \neq \beta$. However, the second order contribution to the variance is order $n_i^{-1}\, n^{-1}$ for $\hat{\mu}_{(2)yi}$ and is order n_i^{-2} for $\hat{\mu}_{(1)yi}$. Also, the regression bias for $\hat{\mu}_{(2)yi}$ is order n^{-1}. If the domains were strata, $\hat{\mu}_{(1)yi}$ might be called the separate regression estimator and $\hat{\mu}_{(2)yi}$ might be called the combined regression estimator.

The estimator $\hat{\mu}_{(3)yi}$ is a synthetic estimator and has a variance of order n^{-1} instead of the order n_i^{-1} variance of the first two estimators. The cost of this reduction in variance is that the bias is order one. Only if the regression line is the same for the domain as for the entire population will the bias be zero.

The average mean square error of the three estimators for any subset of small areas can be estimated. If the n_i are small, the estimated variances will provide only limited information for discriminating among estimators. Likewise, there is only one degree of freedom for bias squared for one particular domain. However, a large domain deviation, relative to the standard error, will lead one to reconsider the synthetic estimator.

In their discussion of models, the authors stress the importance of providing estimators of the reliability for small area estimators. They allude to the fact that the principal estimators of mean square error for model based procedures are estimators of an average mean square error. While this is true, it seems worth mentioning that components-of-variance procedures do not assume the mean square errors to be the same in each domain. Also, for the typical survey situation, the estimators of mean square error need not be constant over domains. For example, one of the terms in the mean square error estimator of the components of variance procedure is the estimator of the variance of the direct estimator. The estimated variance of the direct estimator will be a function of the domain sample size and can also be a function of the direct estimated variance of the direct estimator for that domain. See Battese, Harter, and Fuller (1988), Harville (1976), Prasad and Rao (1990), and Ghosh and Rao (1993).

In their discussion of designs, the authors explain that the variance function is often relatively flat in the vicinity of the optimum allocation to strata. A slight reallocation of sample among strata can markedly increase the efficiency of domain estimators for a relatively small decrease in the efficiency of the overall estimates. The same is true with respect to the combination of direct and synthetic estimators. Thus, if one has a relatively good idea of the variance component associated with small areas, either from a previous study on the same population or from a study on a similar population, and if one is under pressure to produce estimates in a brief time span, then it is reasonable to assign fixed weights to form the linear combination. The loss in efficiency is apt to be modest and the programming required for estimation construction considerably reduced. One estimator in this class, and the one adopted by many practitioners, is the synthetic estimator.

The authors briefly raise the question of internal consistency associated with the construction of small area estimates. As they say, if one uses a data dependent procedure, such as variance components, for each dependent variable, then one produces estimates that are not internally consistent. One option is to use multivariate procedures. See, for example, Fuller and Harter (1987) and Fay (1987). Another procedure suggested by Fuller (1990) is to construct components of variance estimators for a limited subset of variables and then use these estimates as control variables in a regression procedure. The regression procedure produces weights for the individual observations. Once the weights are constructed, any number of output tables can be constructed and all estimates are internally consistent.

It is my observation that the gains made in most practical domain estimation problems come primarily from the wise use of auxiliary information. Thus, effort directed towards obtaining quality auxiliary information is effort well spent. If we are able to find a variable x that is highly correlated with the variable y, then there is less variability remaining to be allocated between area to area variance and sampling variance.

ACKNOWLEDGEMENTS

I thank Jay Breidt for comments.

REFERENCES

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.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley.

FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

FULLER, W.A., and HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384-395.

GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. Unpublished manuscript. Carleton University, Ottawa, Ontario, Canada.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

COMMENT

GRAHAM KALTON[1]

As Singh, Gambino and Mantel (SGM) indicate, there is a growing demand for surveys to provide domain estimates for domains of various sizes and types. This demand is being experienced in many countries throughout the world. In part it may simply reflect a natural growth in the sophistication of survey analysts, who once were content with national estimates and estimates for a few major domains, but who now want to compare and contrast estimates for many different types of domain. In part it results from the needs of policy makers, who require domain information in order to examine how current policies affect different domains, to predict what effects changes in policies might have, and for policy implementation. Information on administrative area domains (*e.g.*, provinces or states, counties, and school districts) is of particular interest for policy purposes (*e.g.*, for identifying low income areas for government support).

In some circumstances the need for domain estimates of adequate precision can be satisfied within the design-based inference framework that is standardly used in the analysis of survey data. This holds for large domains for which the sample sizes are adequate to give the precision required. It can also hold for small domains provided that they are identified in advance, and the sample design is constructed in a way that provides adequate sample sizes. Thus, for example, in the United States, the National Health and Nutrition Examination Survey and the Continuing Survey of Food Intakes by Individuals use differential sampling fractions by age, sex and race/ethnicity and by age/sex and low income status, respectively, in order to provide adequate samples for the domains created by the cross-classifications of these variables. The U.S. Current Population Survey employs differential sampling fractions across the states in order to be able to produce state-level employment estimates. The limitation of this approach is evident when there is a large number of small domains, in which case the sum of the required sample sizes for each domain produces an extremely large overall sample size. This situation occurs often with small administrative districts, such as counties, school districts, and local employment exchanges. In such cases, it may be necessary to discard the standard design-based inference approach in favor of a model-dependent approach that employs a statistical model in the estimation process to borrow strength from data other than that collected in the survey for the given small area. The model-dependent approach may also be required for unplanned small domains, where the need for oversampling had not been foreseen at the design stage.

In response to the demand for small area estimates, a sizeable literature has developed on model-dependent small area estimation methods. Little has, however, been written on the broader issues of small area estimation discussed in the SGM paper, issues that need more attention. Like the authors, I believe that a cautious approach should be adopted to the use of model-dependent small area estimators. I therefore welcome their discussion of methods to make small area estimates within the design-based framework.

From my perspective, the first approach to making small area estimates is to see whether estimates can be produced with adequate precision within the design-based framework. If the domains have been identified in advance, consideration should be given to designing the sample to meet the needs for small area estimates. This may involve ensuring that the small areas do not overlap strata, and ensuring a sufficient sample size for each small area. Another approach suggested by SGM is to minimize the amount of clustering. The smaller the amount of clustering, the less the sample size in each small area is subject to the vagaries of chance. In this regard I see the benefits of less clustering as mainly directed at providing the ability to produce estimates for small areas that were not identified at the design stage. When small areas for which estimates are planned are made into separate strata, the sample size in each small area should be under adequate control even with a clustered sample (provided that the measures of size used in the PPES sampling are reasonable). However, even with planned estimates, there will often be an issue of how to compute variance estimates for a small area from a clustered design, since the number of PSUs sampled in each small area is likely to be small. A variance estimate based on the PSUs within the small area will then be imprecise, with few degrees of freedom, and a generalized variance function approach may be preferred (*e.g.*, assuming that the national design effect applies for each small area). In other words, although the estimate itself may be a design-based estimate, the estimate of its variance may be an indirect one, borrowing strength from other areas. This consideration favors as unclustered a design as possible even for planned small area estimates. The need to model variances is, however, of lesser concern than the need to model the estimates themselves.

An integral part of the design-based framework is a recognition that auxiliary information available for the population may be used at the design stage, at the analysis stage, or at both stages. When information on auxiliary

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

variables that are closely related to the survey variable is available, substantial gains in precision can accrue. The use of auxiliary information at the analysis stage, through such techniques as post-stratification and ratio, regression and difference estimation, has a special appeal for small area estimation. It should be emphasized that ratio and regression estimators may be motivated by assumptions about the model relating the survey variable (Y) and the auxiliary variables (X), but that the resultant estimators are design-consistent irrespective of the appropriateness of the model. The use of an appropriate model produces the greatest gains in precision, but the estimates are approximately unbiased whatever model is chosen. This may be seen in a simple case where variables $X_1, X_2, \ldots, X_p$ are known for every element in the population, and the linear combination $\tilde{Y}_i = B_0 + B_1X_{1i} + \ldots + B_pX_{pi}$ is used to estimate Y_i, the value of the Y-variable for population element i. Assume, for simplicity that the B's are determined from external data, not dependent on the sample. With $Y_i = \tilde{Y}_i + e_i$, the domain total is $Y_a = \sum_{i\epsilon a}\tilde{Y}_i + \sum_{i\epsilon a} e_i = \tilde{Y}_a + E_a$. Since $\tilde{Y}_a$ is known, the estimation problem is one of estimating E_a. From a sample of elements in domain a, E_a may be estimated by $\hat{E}_a = \sum_{j\epsilon s_a} e_j/\pi_j$, where π_j is the selection probability for element j in the sample. The estimator $\hat{E}_a$ is unbiased, independent of the validity of the model employed. The estimation procedure in fact translates the estimation problem from one of estimating Y_a directly to one of estimating E_a and adding on a known constant $\tilde{Y}_a$. To be effective, the procedure requires the domain variance of the e_i to be smaller than that of the Y_i. There is no requirement that $E_a = 0$. The general logic remains the same in the more usual situation where the B's are estimated from the sample. In this case, the estimate of Y_a is design-consistent, irrespective of the model adopted (Särndal 1984). Moreover, the B's may be estimated from the sample data only for the domain of interest, producing what SGM term a direct estimator, or from the total sample, producing a modified direct estimator. A key consideration in the choice between the direct and modified direct estimators in this case is whether the overall B's also apply for the domain. If not, interaction terms between the X's and the domain indicators are called for in the total sample model. With a full set of these interaction terms, the modified direct estimator in effect then reduces to the direct estimator.

The need for a model-dependent approach occurs when the design-based estimate lacks sufficient precision even after the auxiliary data available have been used in as effective a manner as possible. Indeed, in some cases the computation of a direct estimate may be impossible because there are no sample cases in the small area. In such situations, it becomes necessary to use a statistical model to borrow strength from other data, often data from other areas. Such models are built upon assumptions (*e.g.*, $E_a = 0$ in the above example), and the quality of the resultant small area estimates depends on the suitability of the assumptions made. The assumptions are inevitably incorrect to some degree, leading to biases in the small area estimates. Since indirect estimates are biased, the design-based mean square error (MSE) is widely used as the measure of their quality, where MSE $= V' + B^2$ and V' is the variance and B is the bias of the estimate.

The common way to compare the quality of a direct and an indirect estimate is to compare the variance, V, of the former with the MSE of the latter. However, reading the paper caused me to question whether the MSE is the appropriate measure of quality of an indirect estimator. In a practical setting the variance V of the direct estimate can be estimated whereas the design-based MSE of the indirect estimate cannot. In view of this situation, if V = MSE, then the direct estimator would be clearly preferred. In fact, the direct estimator may tend to be preferred if the direct estimator has adequate precision, irrespective of the likely relative magnitudes of V and MSE. In other cases, if B is the expected bias, then the direct estimator may be preferred to the indirect estimator unless $V > V' + kB^2$, where k is a multiplier greater than 1 that allows for the fact that the unknown bias may be larger than expected.

The same argument can be applied to combined (or composite) estimators that employ a weighted average of a direct and an indirect estimator. Often the principle for choosing the weights is taken to be to minimize the mean square error of the combined estimator, leading to weights for the direct and indirect estimators that are inversely proportional to V and MSE, respectively. However, following the above argument, an alternative procedure would be to minimize the weight of the indirect estimator, subject to the condition that the combined estimator is sufficiently accurate. Alternatively, the weights could be determined on some maximum likely value of the MSE, rather than the expected MSE, to reduce the risk of serious bias in the combined estimator.

I do not follow the rationale for the sample size dependent estimators described by SGM in equation (6.11) and (6.12) in general, but under certain assumptions they may be seen to fit in to the logic given above. With an equal probability sample design and $\delta = 1$, these estimators reduce to the direct estimator when the achieved sample size is greater than, or equal to, the expected sample size. If one assumes that the expected sample size gives adequate precision for the small area, this outcome accords with the above reasoning. If the achieved sample size is smaller than expected, the sample size dependent estimator takes a weighted average of a direct and an indirect estimator. If one assumes that the expected sample size is the minimum sample size to give the required precision, this outcome also accords with the above reasoning. If this indeed is the basis of the sample size dependent estimators, then it would seem useful to generalize them to situations where

the expected sample size is not the sample size that just gives the level of precision required.

As has been noted, auxiliary information plays an important role in the production of accurate small area estimates. Such information may be used for improving the precision of design-based estimates or it may be used in the models employed with the model-dependent approach. Ideally auxiliary information that is highly related to the survey variables involved in the estimates is required. The regular compilation of up-to-date auxiliary data for small areas from administrative and other sources can provide a valuable resource for a small area statistics program.

Although the paper mentions the more general problem of small domains, it focuses predominantly on small areas. This is in line with the general literature and the application of indirect estimation procedures. In part, this may be because the number of socio-economic and other small domains of interest (*e.g.*, age/sex domains) is usually relatively small, compared with the numbers of small areas, so that socio-economic domains can be handled by designing the sample to provide direct estimates of adequate precision for each of them. In part, it may be because the definitions of socio-economic and demographic domains are often chosen in the light of the feasibility of producing design-based estimates of adequate precision for them (*e.g.*, using wider age groupings for some domains); in the case of areal domains, however, the areas are predefined, and no collapsing of areas is acceptable. In part, it may be because there is a lack of auxiliary data to use in the statistical models for such domains. In part, it may also be because the analysis of socio-economic domains is often conducted to make comparisons between the domains. Such comparisons are distorted when the estimate for one domain borrows strength from other domains (see, for example, Schaible 1992). This issue brings out the general point that indirect estimates should not be uncritically used for all purposes.

In conclusion, I should like to express my support for the general approach of this paper. Where possible, samples should be designed to produce direct small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates. When indirect estimates are called for, a cautious approach should be used. Models should be developed carefully, estimators that are robust to failures in the model assumptions should be sought, and evaluation studies should be conducted to assess the adequacy of the indirect estimates. Lacking good measures of quality for individual indirect estimates, such estimates need to be clearly distinguished from design-based estimators. Since indirect estimates are not universally valid for all purposes, users need to carefully assess whether the given form of indirect estimate will satisfy their particular needs.

REFERENCES

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

RESPONSE FROM THE AUTHORS

We would like to thank Wayne Fuller and Graham Kalton for their stimulating comments, which we find to be quite complementary to the position developed in our paper. In many cases their comments make certain points clearer and strengthen the arguments presented. Encouraged with this kind of endorsement we would like to carry some of the points about survey design further, while responding to the main points made by the discussants.

There is no doubt that survey designers try to optimize the design under operational constraints to meet the stated objectives of a survey. There are usually several objectives to be met by major surveys and it is quite likely that designers have limited influence in the setting of priorities among the various competing objectives. Nevertheless, it is at this stage of priority setting that the case for small area needs should be made strongly, particularly for major continuing surveys.

During the sixties and seventies emphasis in most countries was placed on sub-national (state/provincial) estimates and certain compromises were made to the earlier designs that optimized national estimates. For example, different sampling fractions were used to ensure a minimum sample size for smaller states/provinces. With the demands for data at the sub-state/province level, such as, county, district and municipality, more compromises to the national optimum allocation become necessary, requiring differing sampling fractions among the administrative areas within states/provinces. For example, if the aim is to produce sub-provincial estimates of comparable quality, then provinces will likely receive sample roughly proportional to the number of subprovincial regions they contain. Such an allocation may not be the same as one using the relative population sizes of the provinces. As we discussed in section 5.4, the allocation approach should put more emphasis on a bottom-up strategy. Losses at higher levels and gains at lower levels would differ from survey to survey but it is likely that in many cases a minor loss in CV at the national level will lead to appreciable gains at small area levels.

Kalton stresses the importance of reduced clustering for variance estimation; it is advantageous to increase the degrees of freedom by having a large number of smaller clusters rather than a small number of larger clusters. We would like to emphasize that clustering has another drawback for estimation, and especially small area estimation, namely, a highly clustered design will lead to high design effects, even for planned small domains. The usual reason for resorting to clustered designs is to reduce survey costs. In light of the changes that continue to occur in the data collection process, such as decreased reliance on at-home interviews and increased use of computer assisted interviewing, a periodic review of the cost-variance models that underlie clustering decisions is necessary.

One other issue not addressed in our paper is the impact of sample rotation in continuous surveys. For a given time point, there may be insufficient sample in some small domains to produce reliable estimates. But, as units rotate out of the sample and are replaced, the accumulated or effective sample in the domains increases and may allow the computation of reliable, albeit time-biased, domain estimates. By judicious choice of rotation schemes, survey designers can maximize the cumulative sample size over some time period. For example, for quarterly estimates in a monthly survey, the optimal rotation pattern is $[1(2)]^k$, *i.e.*, repeat the sequence "one month in sample, two months out" k times. This thinking is in the same spirit as Leslie Kish's ideas on cumulation of samples over time.

Kalton clarifies and elaborates the cautious approach to the use of indirect estimators by suggesting a weighted mean squared error, which attaches a weight greater than 1 to the bias term, to allow for the fact that the bias of the indirect estimator may be larger than expected. There are two distinct reasons why the bias may be larger than what is expected from the model for small area effects: random variation within the model, and model breakdown. It is worth recalling here the suggestion of Fay and Herriot (1979) to constrain a combined estimate to be within one standard error of a design estimate; this approach makes allowance for the possibility of large bias in the model estimator for whatever reason. Kalton also reiterates our position that if a direct estimator is of acceptable quality, then in practice, one may decide to use this direct estimator even though its estimated mean squared error exceeds that of model-based competitors. Because there is always the possibility of model failure lurking in the background, this "better safe than sorry" approach is desirable, at least until some experience with particular indirect estimators in specific situations has been gained. This does not contradict the view that there arise situations in which it is necessary to throw caution to the wind.

In his remarks on the sample size dependent estimator, Kalton's comments imply that there is a risk in the strategy which gives the synthetic component zero weight if the observed sample size in the small domain exceeds the expected sample size there since the latter may be too small to yield adequate direct estimates. One option is to use a value $n_{\min}$ which is the size that produces direct estimates that are just barely acceptable. Note, however, that $n_{\min}$ as defined here is characteristic-dependent.

In his comments, Fuller briefly describes an approach to small area estimation that takes advantage of a variance components model and yet has fixed weights for internal consistency among estimators for different characteristics. Besides internal consistency of small area estimates for different characteristics, a second type of consistency that

is sometimes required is that estimates of totals for the set of small areas within a larger area should add up to the published direct estimate for the larger area. One way to achieve this is to benchmark the small area estimates to the direct estimate for the larger area using, for example, a simple ratio adjustment; however, if the ratio adjustment factors depend on the characteristic then this would destroy the first type of consistency. Both types of consistency could be achieved simultaneously if the direct estimators for the larger area are generalized regression estimators, $\hat{Y}_e + (X - \hat{X}_e)\hat{\beta}$, and the modified direct (Section 6.1 in the paper) estimators $\hat{Y}_{sreg,a} = \hat{Y}_{e,a} + (X_a - \hat{X}_{e,a})\hat{\beta}$ are used for small areas.

As Fuller notes, the average squared bias of an estimator for any subset of small areas can be estimated. Here we would like to stress again that the average bias over a set of small areas is not directly relevant for any particular small area. It is for this reason that we prefer to use, whenever possible, estimators that are approximately design unbiased. When use of a model estimator is unavoidable, serious attempts should be made to find appropriate covariates for which reliable auxiliary information is available in order to minimize the residual bias of the model estimator.

Perhaps due to the obvious timeliness problems associated with census data, neither of the discussants commented on censuses as a source of data for smaller domains. In this context it is worth mentioning that some form of ongoing major post-censal survey replacing or supplementing the decennial census long-form may be considered. Such a strategy, called rolling samples, is described by Kish (1990); a similar approach, called continuous measurement, is described by Alexander (1994). This approach provides a number of options which are worth investigating as potentially cost effective means of producing timely statistics for smaller domains.

Lastly, we would like to stress that the emphasis we put on keeping domain estimation in mind at the design stage, particularly for medium size domains, in no way undermines the important role of models in estimating for very small domains.

We hope that the general direction of the strategy proposed in the paper, supplemented by the fine points brought out by the discussants, particularly the support and cautions summarized by Kalton in his concluding paragraph, will be helpful to survey designers and researchers in finding solutions appropriate to the particular problems they are dealing with.

ADDITIONAL REFERENCES

ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Document for presentation at the annual meeting of the Population Association of America, Miami, Florida, May 5, 1994.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-71.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 23–31
Statistics Canada

Small Domain Estimation for Unequal Probability Survey Designs

D. HOLT and D.J. HOLMES[1]

ABSTRACT

The problem of estimating domain totals and means from sample survey data is common. When the domain is large, the observed sample is generally large enough that direct, design-based estimators are sufficiently accurate. But when the domain is small, the observed sample size is small and direct estimators are inadequate. Small area estimation is a particular case in point and alternative methods such as synthetic estimation or model-based estimators have been developed. The two usual facets of such methods are that information is 'borrowed' from other small domains (or areas) so as to obtain more precise estimators of certain parameters and these are then combined with auxiliary information, such as population means or totals, from each small area in turn to obtain a more precise estimate of the domain (or area) mean or total. This paper describes a case involving unequal probability sampling in which no auxiliary population means or totals are available and borrowing strength from other domains is not allowed and yet simple model-based estimators are developed which appear to offer substantial efficiency gains. The approach is motivated by an application to market research but the methods are more widely applicable.

KEY WORDS: Synthetic estimation; Design-based estimation; Small area estimation; Model-based estimation; Market shares.

1. INTRODUCTION

This paper is concerned with the common problem of estimating domain totals and means from a disproportionately allocated sample survey. Some domains may be large, in which case the achieved sample size may be large too and design-based (or direct) estimators will be satisfactory. Some domains may be small, in which case the achieved sample size may be small too and design-based (or direct) estimators will be too imprecise for practical use. The methods proposed will be motivated through the example of estimating sales, market shares and market penetrations for products in a market research survey. The domains are particular auto manufacturers or models. However, the general approach is applicable to other disproportionately allocated surveys of businesses or institutions.

The problem is analogous to that of using synthetic estimation for small area estimation (Gonzales 1973; Gonzales and Hoza 1978; Platek *et al.* 1987). Synthetic estimation usually depends on two factors: (i) the use of auxiliary variables in conjunction with population means or totals for each small area (or domain) to improve estimates through poststratification or regression estimation, and (ii) the improvement of estimates by pooling data across the small areas (or domains). In our situation no auxiliary population means or totals are available and, since the essential objective is to compare domains (*i.e.*, manufacturers and particular products), the idea of borrowing strength between these is inadmissible. A class of synthetic estimators is proposed which uses neither of these two approaches and yet is preferred to the direct survey estimators. The proposed estimators have a simple structure, an interesting interpretation and can be justified under a set of model assumptions which are testable under the general assumption of non-informative survey design.

2. THE MARKET RESEARCH EXAMPLE

Market researchers often estimate the total volume of sales and market shares for each manufacturer of a particular product. We consider the case of autos purchased for company fleet use in a single year. Estimates of totals and market shares are required for each auto manufacturer and for specific models which are widely purchased for fleet use.

The terms 'fleet' and 'company' are each interpreted widely. A fleet car is taken to mean any auto purchased on a commercial as opposed to a private basis, and used in conjunction with a business in the broadest sense. This includes autos purchased for sales representatives which may be purchased in large numbers. It also includes single purchases of luxury cars for company directors and other senior staff of large companies, as well as purchases by small 'companies' such as groups of doctors, or self-employed people such as shop owners. Thus the population of purchasing companies – termed consumers – includes a large number of small companies that purchase only one or two autos every few years.

[1] D. Holt and D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

In the reference period of one year we define Y_{ki} to be the number of autos of product type k purchased by consumer i. The product type k (the domain) may refer to a specific model of a particular manufacturer, or to all models produced by a manufacturer. Thus, $Y_k = \sum_i Y_{ki}$ is the total number of autos of type k purchased by all consumers. Let Z_i be the total number of autos of any kind purchased by consumer i, and $Z = \sum_i Z_i$ be the total number of auto sales. The market share for product type k is defined as $R_k = Y_k/Z$.

We further define

$$Y'_{ki} = 1 \quad \text{if} \quad Y_{ki} > 0$$
$$= 0 \quad \text{if} \quad Y_{ki} = 0$$

and

$$Z'_i = 1 \quad \text{if} \quad Z_i > 0$$
$$= 0 \quad \text{if} \quad Z_i = 0.$$

Thus, Y'_{ki} and Z'_i are indicator variables for consumers who purchase product type k and at least one auto of any kind, respectively, in the reference period. The number of consumers that purchase product k is thus given by $Y'_k = \sum_i Y'_{ki}$ and the total number of consumers purchasing at least one auto of any kind is given by $Z' = \sum_i Z'_i$. The market penetration for product k, in terms of the proportion of consumers buying a car of any type in the reference period who buy type k, is given by $R'_k = Y'_k/Z'$.

The four parameters Y_k, R_k, Y'_k and R'_k are all legitimate targets of inference in market research and are defined as finite population parameters; namely, domain totals or ratios of domain totals.

3. THE SURVEY DESIGN AND DIRECT ESTIMATORS

The survey design was based upon two mutually exclusive frames and may be regarded as a simple stratified design with ten strata. The first frame was a register (Dun and Bradstreet) of 35,000 companies, stratified into eight strata on the basis of the number of employees and whether the company was classified as 'manufacturing' or 'distributing'. The second frame was a large register of 1.4 million British Telecom business subscribers, stratified into 'private' and 'commercial' numbers. Note that both private and commercial numbers were business subscribers but commercial numbers were allocated if separate commercial premises were occupied.

Using previous survey data the sample was optimally allocated using Neyman allocation to minimize the variance of the estimator of the total number of autos purchased (Z). Data on auto purchases were collected immediately after the end of the reference year. The strata sizes $\{N_h\}$ and sample allocations $\{n_h\}$ for strata $h = 1, \ldots, 10$ are given in Table 1.

Table 1

Sampling Frame: Sample Size and Weight by Stratum

Stratum (h)	Stratum Size N_h	Sample Size n_h	Weight $\pi_h^{-1} = N_h/n_h$
British Telecom:			
Private	389,445	1,150	338.65
Commercial	1,007,399	7,406	136.02
Dun and Bradstreet:			
Manufacturing			
50-99 employees	6,646	235	28.28
100-499	6,826	1,113	6.13
500-999	992	520	1.91
1,000+	1,110	849	1.31
Distributing			
50-99 employees	8,703	472	18.44
100-499	7,625	1,437	5.31
500-999	1,133	484	2.34
1,000+	1,523	1,117	1.36
Overall	1,431,402	14,783	96.83

The sample is a simple, disproportionately allocated stratified design and the direct estimators and their variances are well known. The stratification results in large differences in sampling weights (1.31 to 338.65) and is useful but far from ideal. Many consumers do not purchase any autos at all in the reference year so that each stratum contains a mixture of zero and non-zero responses. For any particular product k the proportion of zero responses in each stratum is obviously larger.

Table 2 contains the direct survey estimates, estimated standard errors (see Holt and Holmes (1993) for derivation), and coefficients of variation for a selection of products from different auto manufacturers. Products A and B represent all models for two major auto manufacturers. Product C is a single model with a substantial share of the fleet market from manufacturer A. The remaining products have small market shares. Products F and G cater for the executive part of the fleet market. The list is incomplete so that the market shares do not sum to one. Also note that the product categories are not mutually exclusive. In general the survey was judged to perform satisfactorily but it was observed over a period of years that estimates for manufacturers or models with small market shares were unstable. This is best seen in terms of the coefficient of variation which is greater than 0.1 for products with small market shares and can be greater than 0.15 or 0.2 in some cases. This instability also affects the estimates of variance as well as the estimates of total sales or market shares of the products.

Table 2
Direct Survey Estimates, Standard Errors and Coefficients of Variation for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{Y}'_k$	Penetration $\hat{R}'_k$	Total $\hat{Y}_k$	Share $\hat{R}_k$
A	59,890	.3843	270,051	.3781
	(2,651)	(.0144)	(35,704)	(.0315)
	(.044)	(.037)	(.132)	(.083)
B	34,282	.2200	153,518	.2149
	(1,960)	(.0117)	(8,653)	(.0131)
	(.057)	(.053)	(.056)	(.061)
C	23,363	.1499	81,381	.1139
	(1,602)	(.0098)	(17,559)	(.0194)
	(.069)	(.065)	(.216)	(.170)
D	13,857	.0889	25,312	.0354
	(1,311)	(.0081)	(2,906)	(.0039)
	(.095)	(.091)	(.115)	(.110)
E	9,025	.0579	24,370	.0341
	(1,146)	(.0072)	(7,336)	(.0101)
	(.127)	(.124)	(.301)	(.296)
F	5,125	.0329	13,724	.0192
	(676)	(.0043)	(2,369)	(.0030)
	(.132)	(.131)	(.173)	(.156)
G	7,518	.0482	11,031	.0154
	(1,015)	(.0064)	(1,456)	(.0022)
	(.135)	(.133)	(.132)	(.143)

Row 1: estimate Row 2: s.e. Row 3: c.v.

4. A MODEL-BASED APPROACH

Given the sample design there is no prospect of improving the efficiency of the direct survey estimators within the conventional sample survey framework. The usual approaches are through the use of auxiliary information for poststratification, ratio or regression estimation but all of these require knowledge of population means or totals. No such information is available. We turn instead to a model-based approach to provide alternative estimators for the whole range of products.

4.1 Estimating Y'_k: the Number of Consumers Purchasing Product Type k

We consider, initially, the number of consumers who buy product type k. We extend the notation from Y'_{ki} to Y'_{khi} in the obvious way to define the indicator random variable of purchase for product k for consumer i in stratum h. We treat each consumer's decision as the outcome of a Bernoulli trial. Let $P_{k|h}$ be the probability that a consumer in stratum h buys an auto of type k [$P_{k|h}$ = Prob$(Y'_{khi} = 1)$]. We define the model-based equivalent of Y'_k, the total number of consumers of product k, as

$$\Theta'_k = \sum_h N_h P_{k|h}. \tag{1}$$

Assuming that each consumer's decision is independent the likelihood may be written as the usual product of binomial terms. The maximum likelihood estimators are given by $\hat{P}_{k|h} = n_{kh}/n_h$, and the maximum likelihood estimator of Θ'_k is the familiar stratified sampling estimator

$$\hat{\Theta}'_k(1) = \sum_h \frac{N_h}{n_h} n_{kh} = \sum_h N_h \bar{y}'_{kh}, \tag{2}$$

where n_{kh} is the sample count of consumers in stratum h that buy product k, n_h is the stratum sample size and $\bar{y}'_{kh} = n_{kh}/n_h$ is the sample mean for consumers in stratum h (*i.e.*, the sample proportion of consumers in stratum h who buy product k). This estimator is generally unsatisfactory when the sample size for product k is too small.

Suppose we introduce an additional conditioning factor such that every consumer may be categorized into one of its categories f, $f = 1, \ldots, F$, and further extend the definition of the indicator random variable to Y'_{khfi}. These categories f will cut across the strata h and the idea is to define f so that, within any particular category, whether a consumer buys product type k or not is independent of the stratum membership h. In the case of fleet purchases we define a categorization based on the total number of autos owned and operated by each consumer (*i.e.*, the fleet size). A more detailed discussion of the choice of f is given in Section 5.

If N_{hf}, the population counts of consumers in stratum h and fleet size category f, are known then (1) may be extended in the obvious way and the target parameter can now be expressed as

$$\Theta'_k = \sum_h \sum_f N_{hf} P_{k|hf}. \tag{3}$$

Equation (3) is the case of poststratification if $\{N_{hf}\}$ are known, and in this case the additional information will lead to a gain in efficiency (Holt and Smith 1979). When $\{N_{hf}\}$ are unknown we may rewrite the model in terms of two sets of probabilities:

$Q_{f|h}$ = Prob {consumer has fleet size f | stratum h},

$P_{k|hf}$ = Prob {consumer buys product type k | stratum h and fleet size f}.

The target parameter may now be expressed as

$$\Theta'_k = \sum_h \sum_f N_h Q_{f|h} P_{k|hf}.$$

To obtain an alternative model-based estimator we make further assumptions about the model parameters. Suppose now that

$$P_{k|hf} = P_{k|f} \quad \text{for all} \quad h. \tag{5}$$

This implies that conditional on the categorization f (the size of the fleet operated by a consumer), the probability of buying product type k is *independent* of the original stratum membership h. Algebraically, the assumption is analogous to that used in synthetic estimation for small area estimation but in that case information is pooled across areas. That form of the assumption is inadmissible in our case. We choose instead pooling across strata within the domain of study. The idea is to choose a conditioning variable which accounts for the marginal association between choice of product and stratum membership.

Using assumption (5) and with the obvious extension of the notation ($n_{kf} = \sum_h n_{khf}$, *etc.*) it may be shown that

$$\hat{Q}_{f|h} = \frac{n_{hf}}{n_h}, \qquad \hat{P}_{k|f} = \frac{n_{kf}}{n_f}$$

and the maximum likelihood estimator of Θ'_k becomes

$$\hat{\Theta}'_k(2) = \sum_h \sum_f N_h \frac{n_{hf}}{n_h} \frac{n_{kf}}{n_f} = \sum_f \hat{N}_f \frac{n_{kf}}{n_f}$$

$$= \sum_h \hat{N}_f \bar{y}'_{kf}, \tag{6}$$

where $\hat{N}_f = \sum_h N_h \, n_{hf}/n_h$, and $\bar{y}'_{kf} = n_{kf}/n_f$ is the unweighted sample mean for consumers in category f (*i.e.* the sample proportion of consumers in category f who buy product k).

Thus (6) has the form of a stratified estimator based on the categorization f but with the population sizes in each stratum $\{N_f\}$ unknown. Note that an estimator of this form, but with known $\{N_f\}$, would arise naturally if a stratified sample based on f had been selected. In fact this is **not** so: the sample members of category f are **not** selected with equal probability. However, the parameter assumptions lead to treating the sample in each category f as if it was an equal probability sample since under assumption (5) the sample weights are uninformative and simply lead to efficiency loss when estimating $P_{k|f}$. Hence, although the sampling fractions n_h/N_h are used to estimate $\{N_f\}$ they are not used explicitly in $\hat{P}_{k|f} = n_{kf}/n_f = \bar{y}'_{kf}$. Note that the estimator pools information across strata h, within domain k but **not** between domains (*i.e.* products).

Note that if n_h/N_h is constant, equation (6) reduces to the usual expansion estimator given by (2), and assumption (5) has not yielded a new estimator. If the sample is disproportionately allocated the assumption leads to the use of the sampling weights for $\hat{N}_f$ (where they are needed) but not for estimating $P_{k|f}$ (where they are uninformative given f and assumption (5)).

Equation (5) is a strong set of assumptions, requiring $P_{k|hf}$ to be exactly equal to a common value $P_{k|f}$ for all h. In practice, random assumptions such as $P_{k|hf} = P_{k|f} + \epsilon_{k|hf}$ may be introduced, where $E[\epsilon_{k|hf}] = 0$ and $V[\epsilon_{k|hf}] = \sigma^2_\epsilon$. These assumptions will lead to hierarchical Bayes or empirical Bayes analysis as described in Ghosh and Rao (1994) or Fay and Herriot (1979). These methods are not developed here since the simple form of the model-based estimator would be lost, together with the insight that this provides. In a similar vein the approach of Särndal and Hidiriglou (1989) or Drew, Singh and Choudhry (1982) may be applied to yield sample size dependent estimators without violating the requirement that no information is pooled across domains (products).

We can compare the estimators in (2) and (6) when assumption (5) holds since it may be shown that

$$V_\xi(\hat{\Theta}'_k(1)) = \sum_h \frac{N_h^2}{n_h} P_{k|h}(1 - P_{k|h})$$

$$= \sum_h \sum_f \frac{N_h^2}{n_h} Q_{f|h} P_{k|f}$$

$$- \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} Q_{f|h} Q_{f'|h} P_{k|f} P_{k|f'}, \tag{7}$$

where the notation $V_\xi(\cdot)$ is used to emphasize that the variance is evaluated with respect to the model-based distribution.

It may also be shown that under assumption (5)

$$V_\xi(\hat{\Theta}'_k(2)) = \sum_h \sum_f \frac{N_h^2}{n_h} P_{k|f}^2 \, Q_{f|h} \, (1 - Q_{f|h})$$

$$- \sum_h \sum_{\substack{f \\ f \neq f'}} \sum_{f'} \frac{N_h^2}{n_h} P_{k|f} \, P_{k|f'} \, Q_{f|h} \, Q_{f'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{P_{k|f}(1 - P_{k|f}) Q_{f|h}}{\sum_h n_h \, Q_{f|h}}$$

$$\left\{ (1 - Q_{f|h}) + n_h \, Q_{f|h} \right.$$

$$\left. + \frac{[1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q^2_{f|h}]}{\sum_h n_h \, Q_{f|h}} \right\} \tag{8}$$

and that $V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2)) \geq 0$.

Thus under the additional model assumptions $\hat{\Theta}_k'(2)$ has smaller variance as would be expected. These expressions are model-based variances and no finite population corrections arise. A predictive approach to the unobserved elements in each poststratum would give rise to finite population correction factors.

The maximum likelihood estimator of the market penetration for product type k, R_k', under assumption (5) is simply given by

$$\hat{\Omega}_k'(2) = \frac{\sum_f \hat{N}_f \frac{n_{kf}}{n_f}}{\sum_f \hat{N}_f \frac{n_{af}}{n_f}} = \frac{\sum_f \hat{N}_f \bar{y}_{kf}'}{\sum_f \hat{N}_f \bar{z}_f'}, \qquad (9)$$

where n_{af} is the sample count of consumers in fleet category f that buy an auto of any kind, and $\bar{z}_f' = n_{af}/n_f$ is the sample proportion of consumers in category f who buy an auto of any kind.

4.2 Efficiency of the Model-Based Estimator of Y_k'

To investigate the gain in efficiency of $\hat{\Theta}_k'(2)$ over $\hat{\Theta}_k'(1)$ we consider the efficiency of the model-based estimator, defined by

$$e[\hat{\Theta}_k'(2)] = \frac{V_\xi(\hat{\Theta}_k'(1)) - V_\xi(\hat{\Theta}_k'(2))}{V_\xi(\hat{\Theta}_k'(1))}, \qquad (10)$$

for various population structures in which assumption (5) holds.

We consider a population with strata $\{h\}$, stratum sizes $\{N_h\}$ and sample allocations $\{n_h\}$ as given in Table 1, and a conditioning factor with ten categories f ($f = 1, \ldots, 10$) of increasing fleet size. We compute the efficiency factor $e[\hat{\Theta}_k'(2)]$ for various combinations of parameter values of $\{Q_{f|h}\}$ and $\{P_{k|f}\}$.

We consider five different structures for $\{Q_{f|h}\}$:

(a) $Q_{f|h} = \begin{cases} 1 & f = h \\ 0 & f \neq h \end{cases}$ for $h = 1, \ldots, 10$.

(b) $Q_{f|h} = \begin{cases} 0.95 & f = h \quad \text{for } h = 1, \ldots, 10 \\ 0.025 & f = h - 1 \quad \text{for } h = 2, \ldots, 10 \\ 0.025 & f = h + 1 \quad \text{for } h = 1, \ldots, 9 \\ 0.05 & h = 1, f = 2 \text{ and } h = 10, f = 9 \\ 0 & \text{otherwise} \end{cases}$

= Band Matrix (0.025, 0.95, 0.025).

(c) $Q_{f|h}$ = Band Matrix (0.05, 0.90, 0.05).

(d) $Q_{f|h}$ = Band Matrix (0.05, 0.10, 0.70, 0.10, 0.05).

(e) $Q_{f|h} = 0.1$ for $h = 1, \ldots, 10$ and $f = 1, \ldots, 10$.

We consider four different structures for $\{P_{k|f}\}$:

(i) $P_{k|f} = \begin{cases} 0.1 & f = 1, 2 \\ 0 & \text{otherwise.} \end{cases}$

(ii) $P_{k|f} = 0.1 - 0.01\,(f - 1)$ for $f = 1, \ldots, 10$.

(iii) $P_{k|f} = 0.1f$ for $f = 1, \ldots, 10$.

(iv) $P_{k|f} = 0.5$ for $f = 1, \ldots, 10$.

Structure (a) is one where the categorization f coincides with the stratification. In structures (b), (c) and (d), in any particular stratum h the majority of consumers fall into one fleet category ($f = h$) with a few consumers in neighbouring categories (*e.g.*, for (b) and (c) $f = h - 1$, $h + 1$). Finally, structure (e) implies that, in any stratum h, consumers will be equally likely to fall into any one of the fleet categories $f = 1, \ldots, 10$.

Structure (i) for $P_{k|f}$ implies a type of auto that is purchased with a small probability by consumers with small fleet sizes (*i.e.* that fall in categories $f = 1$ or 2), but not purchased by consumers with large (r) fleet sizes. Structure (ii) suggests a type of auto purchased with small probability which decreases as fleet size increases, whilst structure (iii) implies the reverse. In structure (iv) a popular model is bought with probability 0.5 *regardless* of the consumer's fleet size.

Table 3 gives the efficiency factor defined in (10) for each combination of structures for $Q_{f|h}$ and $P_{k|f}$ under the disproportionate allocation given in Table 1. Column (a) of the table is the special case where the stratification and the categorization f coincide, and the two estimators $\hat{\Theta}_k'(1)$ and $\hat{\Theta}_k'(2)$ are the same. The table shows that large gains in efficiency (*e.g.*, 70%) can be attained for certain parameter combinations: the weaker the association

Table 3
Efficiency Factors, $e[\hat{\Theta}_k'(2)]$, for Various Combinations of $Q_{f|h}$ and $P_{k|f}$

		Structure for $Q_{f\|h}$				
		(a)	(b)	(c)	(d)	(e)
Structure for $P_{k\|f}$	(i)	0	0.108	0.196	0.355	0.648
	(ii)	0	0.116	0.206	0.391	0.695
	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

between f and h the greater the efficiency gain. Even for structures (c) and (d) where the association between f and h is strong, substantial efficiency gains can be achieved. The structure $Q_{f|h}$ is much more important than $P_{k|f}$ in determining efficiency gain.

In the special case (e) where $Q_{f|h}$ is a constant for all f and h it can be shown that the efficiency factor can be expressed as

$$e[\hat{\Theta}_k'(2)] = \left(1 - \frac{\delta^2}{\bar{P}_{k|f}\,(1 - \bar{P}_{k|f})}\right) \frac{\sum_h \tau_h N_h^2/n_h}{\sum_h N_h^2/n_h}, \quad (11)$$

where

$$\bar{P}_{k|f} = \frac{1}{F}\sum_{f=1}^{F} P_{k|f} \quad \text{and} \quad \delta^2 = \frac{1}{F}\sum_{f=1}^{F} (P_{k|f} - \bar{P}_{k|f})^2$$

are the mean and variance of $\{P_{k|f}\}$ over the categories f, and $\tau_h = 1 - n_h/n + O(n^{-1})$. The term in parentheses in (11) lies between 0 and 1 and it's value depends on how the $\{P_{k|f}\}$ vary over the categories f. In case (iv) $P_{k|f}$ is constant and so this term is unity. The second term of (11) depends solely on the design, and its value for the sample allocation specified in Table 1 is 0.706.

4.3 Estimating Y_k: the Number of Autos Purchased of Product Type k

The previous approach in Section 4.1 may be extended to the number of purchases. We introduce a further conditioning factor which represents the total number of autos purchased, m, regardless of product type, and we extend the notation in the obvious manner to Y_{khfmi}, the random variable representing the number of autos of product type k purchased by consumer i in stratum h, fleet size f, and buying m autos of any kind. The idea is that the number of purchases of product k is likely to vary depending on the total number of autos purchased. Let

$S_{m|hf}$ = Prob{consumer buys m autos of any kind | h,f}, $m = 0, 1, 2, \ldots,$

$T_{\ell|hfm}$ = Prob{consumer buys ℓ autos of type k | h,f,m}, $\ell = 0, 1, \ldots, m.$

The model-based target parameter, equivalent to the total purchases of product k, Y_k, is extended from (4) and may now be expressed as

$$\Theta_k = \sum_h \sum_f \sum_m \sum_\ell N_h Q_{f|h} S_{m|hf} T_{\ell|hfm}\, \ell. \quad (12)$$

We consider two sets of additional assumptions, the first of which is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h. \quad (13)$$

These assumptions imply that conditional on fleet size category, f, and the total number of new autos purchased, m, the distribution of the number of autos purchased of product type k is independent of stratum h.

The maximum likelihood estimator of Θ_k under assumptions (13) is

$$\hat{\Theta}_k(2) = \sum_f \sum_m \hat{N}_{fm}\, \bar{y}_{kfm}, \quad (14)$$

where $\hat{N}_{fm} = \sum_h N_h n_{hfm}/n_h$, and $\bar{y}_{kfm} = \sum_\ell \ell\, n_{fm\ell}/n_{fm}$ is the unweighted sample mean of the number of autos of product type k purchased by consumers of fleet size f that purchased a total of m autos of any kind.

The selection probabilities are used here to provide a weighted estimator of N_{fm}, the total number of consumers of fleet size f that buy m cars of any kind. The form of the estimator is analogous to that in equation (6). Under the model assumption (13) it may be shown that

$$\begin{aligned} V_\xi(\hat{\Theta}_k(2)) = {} & \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \mu_{fm}^2 Q_{fm|h}\,(1 - Q_{fm|h}) \\ & - \sum_h \sum_f \sum_m \sum_{f'} \sum_{m'} \frac{N_h^2}{n_h} \mu_{fm}\mu_{f'm'} Q_{fm|h} Q_{f'm'|h} \\ & \qquad (f,m) \neq (f',m') \\ & + \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \frac{\sigma_{fm}^2 Q_{fm|h}}{\sum_h n_h Q_{fm|h}} \\ & \left\{ (1 - Q_{fm|h}) + n_h Q_{fm|h} \right. \\ & \left. + \frac{[1 + (2n_h - 3)Q_{fm|h} - 2(n_h - 1)Q_{fm|h}^2]}{\sum_h n_h Q_{fm|h}} \right\}, \end{aligned} \quad (15)$$

where $Q_{fm|h} = Q_{f|h}\, S_{m|hf}$, $\mu_{fm} = E_\xi\{Y_{khfmi}\}$, and $\sigma_{fm}^2 = V_\xi\{Y_{khfmi}\}$.

In practice, $\bar{y}_{kfm}$ will be based on very few observations if few customers in fleet size category f purchase exactly m cars. For more stability m may be defined as an ordinal variable by grouping the total number of autos purchased into a small number of categories. In this case assumption (13) implies that the distribution of purchases for product type k is the same within fleet size category f and total

purchase category m. Also, ℓ may be treated as a continuous random variable and distributional assumptions made about ℓ leading to ratio or regression estimators.

A second and even stronger set of parameter assumptions is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h,$$

$$S_{m|hf} = S_{m|f} \quad \text{for all} \quad h. \tag{16}$$

These assumptions imply that conditional on fleet size category, f, the joint distribution of the number of autos purchased of type k and the total number of autos purchased of any kind, m, is independent of the stratum h. In this case the maximum likelihood estimator of Θ_k is given by

$$\hat{\Theta}_k(3) = \sum_f \hat{N}_f \bar{y}_{kf}, \tag{17}$$

where $\bar{y}_{kf} = \sum_\ell \ell\, n_{f\ell}/n_f$ is the unweighted sample mean of the number of autos of product type k purchased by consumers in fleet size f regardless of how many autos the consumer bought in total, and $\hat{N}_f = \sum_h N_h\, n_{hf}/n_h$ is a weighted estimator of the number of consumers of fleet size f overall. It may be shown that under assumptions (16)

$$\begin{aligned} V_\xi(\hat{\Theta}_k(3)) = {} & \sum_h \sum_f \frac{N_h^2}{n_h} \mu_f^2 Q_{f|h} (1 - Q_{f|h}) \\ & - \sum_h \sum_{\substack{f \\ f \neq f'}} \sum_{f'} \frac{N_h^2}{n_h} \mu_f \mu_{f'} Q_{f|h} Q_{f'|h} \\ & + \sum_h \sum_f \frac{N_h^2}{n_h} \frac{\sigma_f^2 Q_{f|h}}{\sum_h n_h Q_{f|h}} \\ & \left\{ (1 - Q_{f|h}) + n_h Q_{f|h} \right. \\ & \left. + \left[\frac{1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2}{\sum_h n_h Q_{f|h}} \right] \right\}. \end{aligned} \tag{18}$$

If assumptions (16) were plausible then $\bar{y}_{kf}$ would be based on larger sample sizes than $\bar{y}_{kfm}$ in (14) and hence $\hat{\Theta}_k(3)$ would be more stable.

The maximum likelihood estimator of the market share for product type k, R_k, under assumption (16), is given by

$$\hat{\Omega}_k(3) = \frac{\sum_f \hat{N}_f \bar{y}_{kf}}{\sum_f \hat{N}_f \bar{z}_f}, \tag{19}$$

where $\bar{z}_f$, defined analogously to $\bar{y}_{kf}$, is the unweighted sample mean number of autos of any kind purchased by consumers in fleet category f.

5. EMPIRICAL RESULTS

5.1 Estimating Consumers

In Section 4.2 the efficiency of $\hat{\Theta}_k'(2)$ was investigated for various population structures when assumption (5) held. Readers may find this measure unconvincing since (5) will not hold in practice. We now use the actual survey data to compute $\hat{\Theta}_k'(2)$ for a particular categorization of the conditioning factor that is defined by a combination of the fleet size *and* whether or not the consumer purchased any autos of any kind for fleet use (see Table 4). Empirical evaluations of synthetic estimators have been carried out by Schaible, Brock and Schnack (1977) and Drew, Singh and Choudhry (1982) in different contexts.

For each of the products A-G listed in Table 2 a χ^2 test was used to test the hypothesis that, conditional on the category of the conditioning factor (f), whether or not a consumer purchases that product is independent of stratum (h). Note that for our example the design is stratified random sampling and standard multinomial assumptions apply. For multistage designs, the standard χ^2 analysis would have to be adjusted by using Rao-Scott adjustments for example. In practice it is difficult to find a categorization f such that conditional independence assumptions (5) hold for every product type. However, for the categorization defined in Table 4 it was found that

Table 4

Definition of the Categories, f, of the Conditioning Factor

Categories f	Definition of f	
	Fleet Size	Fleet Purchases
1	Any	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

most of the variability in the probability of purchasing a particular product type was explained by the category f of the conditioning factor and very little of the residual variation was due to differences in strata.

The model-based estimates for consumers, $\hat{\Theta}_k'(2)$ and $\hat{\Omega}_k'(2)$, obtained from (6) and (9) respectively, are given in Table 5. The model-based variances may give an optimistic view of the precision of the estimators since they depend on the conditional independence assumptions in the model which may be untrue in practice. Alternatively the usual survey estimate of the p-based variance of the model-based estimator may be derived (see Holt and Holmes 1993). This requires no distributional or conditional independence assumptions of any kind and might be considered a more objective measure. These estimates of standard errors are given in Table 5. Since the estimated standard errors are design-based, they include finite population corrections. [We note here that the model-based standard errors for $\hat{\Theta}_k'(2)$ (not shown in Table 5) were consistently around 10% smaller than the p-based standard errors].

Table 5

Model-Based Estimates with p-Based Standard Errors for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{\Theta}_k'(2)$	Penetration $\hat{\Omega}_k'(2)$	Total $\hat{\Theta}_k(3)$	Share $\hat{\Omega}_k(3)$
A	63,433	.4070	263,511	.3722
	(2,230)	(.0105)	(13,007)	(.0048)
B	39,673	.2546	177,067	.2501
	(1,587)	(.0086)	(9,530)	(.0046)
C	21,930	.1407	65,357	.0923
	(1,142)	(.0066)	(3,836)	(.0027)
D	13,422	.0861	22,146	.0313
	(868)	(.0052)	(1,351)	(.0016)
E	7,366	.0473	15,798	.0223
	(675)	(.0041)	(1,223)	(.0014)
F	5,826	.0374	14,398	.0203
	(492)	(.0031)	(1,113)	(.0012)
G	7,686	.0493	11,207	.0158
	(633)	(.0039)	(813)	(.0011)

Row 1: estimate Row 2: p-based s.e.

Comparing these results with the usual survey results given in Table 2 we find that the standard errors for estimating totals are considerably smaller – around 30-40% smaller for all products except A and B (the major manufacturers) where the reduction is about 15-20%. This pattern is expected since the original survey design was optimal for the total sales of autos and therefore relatively efficient for products with a large market share. We expect the products with smaller market shares to benefit most from the model-based approach.

For estimating market penetration the reduction in standard error is again about 30-40% with slightly smaller reductions for products A and B.

5.2 Estimating Autos

Table 5 also contains model-based estimates for the total number of autos purchased of type k and the corresponding market share, $\hat{\Theta}_k(3)$ and $\hat{\Omega}_k(3)$ as defined by (17) and (19) respectively, for the *same* categorization f of the conditioning factor as given in Table 4. P-based standard errors for these estimates are also presented in Table 5.

Comparing with the standard survey estimates given in Table 2 large reductions in standard errors for estimating totals are obtained (40-80%) apart from product type B. Similarly, for estimating the market shares the reduction in standard error is again substantial.

6. DISCUSSION

The model-based estimators are derived using conditional independence assumptions to partition the estimation problem into two components. The first, an estimate of N_f (the number of consumers of fleet size f), makes use of the unequal selection probabilities, whereas the second, an estimate of the proportion of consumers of fleet size f buying product type k (or the average number of autos of product type k purchased by consumers of fleet size f) does not. This can result in a substantial efficiency gain.

If the conditional independence assumptions are invalid then in ordinary design-based terms the estimators will have a residual bias but this may be an acceptable risk to achieve stability of the estimators over the whole product range. For the numerical results in previous sections, only the model-based estimates for product B are outside of the 95% confidence interval based on the direct survey estimator. The conditional independence assumptions will depend on the choice of the categories f, and can be tested using chi-square tests for contingency tables.

Whilst the results in Table 5 show that the design-based standard errors for the model-based estimates are generally smaller than for the direct estimates shown in Table 2, it may be argued that the model-based estimators may be biased and hence provide no gain in terms of mean-squared error (MSE). The bias will arise from the inappropriateness of the conditional independence assumptions (*e.g.*, equation (5)). This is not testable, but a comparison of Tables 2 and 5 can give some insight into the size of bias that would be required to cause the MSE to be the same

for both the direct and the model-based estimators. Consider the estimate of total consumers for product E which is strongly affected by the procedure and hence perhaps most susceptible to bias. The variance (and hence MSE) of the direct estimator is $1,146^2 = 1,313,316$ whereas for the model-based estimator the variance is $675^2 = 455,625$. Hence, the model-based estimate of 7,366 would need a bias of 926 in order for the MSEs to be the same.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 19-47.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. To appear in *Statistical Science*.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *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.

HOLT, D., and HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.

HOLT, D., and SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Ser. A, 142, 33-46.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.

SÄRNDAL, C.-E., and HIDIRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

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.

Survey Methodology, June 1994
Vol. 20, No. 1, pp. 33–43
Statistics Canada

Time Series EBLUPs for Small Areas Using Survey Data

A.C. SINGH, H.J. MANTEL and B.W. THOMAS[1]

ABSTRACT

In estimation for small areas it is common to borrow strength from other small areas since the direct survey estimates often have large sampling variability. A class of methods called composite estimation addresses the problem by using a linear combination of direct and synthetic estimators. The synthetic component is based on a model which connects small area means cross-sectionally (over areas) and/or over time. A cross-sectional empirical best linear unbiased predictor (EBLUP) is a composite estimator based on a linear regression model with small area effects. In this paper we consider three models to generalize the cross-sectional EBLUP to use data from more than one time point. In the first model, regression parameters are random and serially dependent but the small area effects are assumed to be independent over time. In the second model, regression parameters are nonrandom and may take common values over time but the small area effects are serially dependent. The third model is more general in that regression parameters and small area effects are assumed to be serially dependent. The resulting estimators, as well as some cross-sectional estimators, are evaluated using bi-annual data from Statistics Canada's National Farm Survey and January Farm Survey.

KEY WORDS: Composite estimation; State space models; Kalman filter; Fay-Herriot estimator.

1. INTRODUCTION

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). Small area estimation techniques in use in U.S. federal statistical programs are reviewed by the Federal Committee on Statistical Methodology (1993). The basic idea underlying all small area methods is to borrow strength from other areas by assuming that different areas are linked via a model containing auxiliary variables from the supplementary data. It would also be important to borrow strength across time because many surveys are repeated over time. Recently time series methods have been employed to develop improved estimators for small areas; see Pfeffermann and Burck (1990) and Rao and Yu (1992). It is interesting to note that after the initiative of Scott and Smith (1974) on the application of time series methods to survey data, there has only lately been a resurgence of interest in developing suitable estimates of aggregates from complex surveys repeated at regular time intervals; see *e.g.*, Bell and Hillmer (1987), Binder and Dick (1989), Pfeffermann (1991), and Tiller (1992).

In this paper we consider some natural generalizations of the best linear unbiased predictor (BLUP) for small areas when a time series of direct small area estimates is available. An important example of the BLUP for small areas is the Fay-Herriot (FH) estimator, which entails smoothing of direct estimators by cross-sectional modelling of small area totals. The resulting estimators are composite estimators (*i.e.*, convex combinations of direct and synthetic estimators) and are called empirical BLUPs, or EBLUPs, whenever estimates of some variance components are substituted in the BLUPs. The work of Fay and Herriot (1979) represents an important milestone in the field of small area estimation because it is probably the first example of a large scale application of small area estimation by government agencies for policy analysis. With the use of structural models, we derive time series EBLUPs which combine both cross-sectional and time series data. The models underlying the time series EBLUPs were chosen on the basis of general heuristic considerations rather than formal model testing procedures. Formal testing of these types of models with survey data is very difficult and not very much is available. Instead, we begin with a regression model that is reasonable for the larger area, and then allow random small area effects to account for any local deviations from the global model. The regression parameters and random small area effects are allowed to evolve over time according to a state space model that was also formulated heuristically. We have not considered here the problem of mean squared error (MSE) estimation for our estimators. MSEs with respect to the motivating models could be defined and estimated for many of the estimators; however, the focus of this paper is on the performance of the estimators in a repeated sampling framework. MSE estimation is an important and difficult problem, and the availability of reliable MSE estimators could be an important consideration in the choice of estimators.

[1] A.C. Singh and H.J. Mantel, Social Survey Methods Division; B.W. Thomas, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The main purpose of this paper is to compare time series EBLUPs with cross-sectional estimators such as post-stratified domain, synthetic, FH and sample size dependent estimators. In the time series modelling of the direct small area estimates we assume that the survey errors are uncorrelated over time. When survey errors are correlated over time and can be modelled reasonably (*e.g.*, ARMA) the approach of Pfeffermann (1991) can be used to obtain time series EBLUPs via the Kalman filter. Rao and Yu (1992) obtain EBLUPs for a model, in which the Kalman filter cannot be applied, with survey errors having arbitrary correlation structure over time but being uncorrelated across areas. They also develop second order approximations to, and estimation of, the mean squared error under their model. When a model for the correlated survey errors is difficult to specify it may be possible, using a suitably modified Kalman filter, to get good sub-optimal estimators (Singh and Mantel 1991).

In this paper we report on an empirical study of the efficiency of time series EBLUPs. The study uses Monte Carlo simulations from real time series data obtained from Statistics Canada's biannual farm surveys. The main findings of the study are

(i) There can be reasonable gains in efficiency with time series EBLUPs over cross-sectional estimators.

(ii) Within the class of time series methods considered in this paper, introduction of serial dependence in the random small area effects is found to be beneficial.

(iii) Although any smoothed version of the direct small area estimator is expected to be biased, the time series EBLUPs exhibit less bias than cross-sectional smoothing methods.

Section 2 contains a description of various cross-sectional methods for small area estimation. Time series EBLUPs are described in Section 3 and the details and results of the Monte Carlo comparative study are given in Section 4. Finally, Section 5 contains concluding remarks.

2. METHODS BASED ON CROSS-SECTIONAL DATA

In this section we describe some well known small area estimation methods that use survey data from only the current time. Ghosh and Rao (1994) contains a good survey of various small area estimators.

Let $\underline{\Theta}$ denote the vector of small area population totals Θ_k, $k = 1, \ldots, K$. In this section, which deals with methods based on cross-sectional data, we ignore the dependence of $\underline{\Theta}$ on time t for simplicity.

2.1 Method 1 (Expansion Estimator for Domains)

This estimator is given by

$$g_{1k} = \sum_{j \in s_k} d_j y_j,$$

where d_j is the survey weight for sample unit j. For stratified simple random sampling, which is used for our simulation study in Section 4, we have

$$g_{1k} = \sum_h (N_h/n_h) \sum_{j \in s_{hk}} y_{hj}, \qquad (2.1)$$

where y_{hj} is the j-th observation in the h-th stratum, s_{hk} denotes the set of n_{hk} sample units falling in the k-th small area in the h-th stratum and n_h, N_h denote respectively the sample and population sizes for the h-th stratum. This estimator is often unreliable because n_{hk}, the random sample size in the small area, may be small in expectation and could have high variability. Conditional on the realized sample size n_{hk}, g_{1k} is biased. However, unconditionally, it is unbiased for Θ_k.

2.2 Method 2 (Post-stratified Domain Estimator)

We will also refer to this estimator as the direct small area estimator. If the population size N_{lk} is known for some post-strata indexed by l, then the efficiency of the estimator g_{1k} could be improved by post-stratification. We define

$$g_{2k} = \sum_l N_{lk} \sum_{j \in s_{lk}} d_j y_j \Big/ \sum_{j \in s_{lk}} d_j = \sum_l N_{lk} \bar{y}_{lk}.$$

In our simulations our post-strata are the intersections of design strata with small areas which leads to

$$g_{2k} = \sum_h (N_{hk}/n_{hk}) \sum_{j \in s_{hk}} y_{hj} = \sum_h N_{hk} \bar{y}_{hk}. \qquad (2.2)$$

This estimator also may not be sufficiently reliable because of the possibility of n_{hk}'s being small in expectation. If $n_{hk} = 0$, the above estimator is not defined. It is conventional to replace $\bar{y}_{hk}$ by 0 when $n_{hk} = 0$. In the empirical study presented in this paper, we replaced $\bar{y}_{hk}$ by the synthetic estimate $(\bar{X}_{hk}/\bar{X}_h)\bar{y}_h$, where X is a suitable covariable, whenever $n_{hk} = 0$.

The estimator g_{2k} in (2.2) is conditionally (given $n_{hk} > 0$) unbiased and approximately unconditionally unbiased. Appendix A.1 gives details of estimation of the conditional mean squared error, v_k, of g_{2k}.

2.3 Method 3 (Synthetic Estimator)

It is possible to define a more efficient estimator by assuming a model which allows for "borrowing strength" from other small areas. This gives rise to synthetic estimators, see *e.g.*, Gonzalez (1973) and Ericksen (1974). Suppose different small area totals are connected via the auxiliary variable X_k by a linear model as

$$\Theta_k = \beta_1 + \beta_2 X_k, \; k = 1, \ldots, K, \qquad (2.3a)$$

or in matrix notation

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta}, \tag{2.3b}$$

where $F = (F_1, F_2, \ldots, F_K)'$, $F_k = (1, X_k)'$. Now consider a model for the direct small area estimators g_{2k}'s as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

where $\underset{\sim}{g}_2 = (g_{21}, \ldots, g_{2K})'$, $\underset{\sim}{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)'$, ϵ_ks are uncorrelated survey errors with mean 0 and variance v_k. Note that the g_{2k}s are uncorrelated over areas since they are conditionally (given n_{hk}) unbiased and the samples in different small areas are conditionally independent.

Denoting by $\hat{\underset{\sim}{\beta}}$ the weighted least squares (WLS) estimate of $\underset{\sim}{\beta}$, we obtain the regression-synthetic estimator of Θ_k under the assumed model as

$$\underset{\sim}{g}_3 = F\hat{\underset{\sim}{\beta}}.$$

The above estimator could be heavily biased unless the model (2.3) is satisfied reasonably well. The above model may not be realistic because no random fluctuation or random small area effect (a_k, say) is allowed.

2.4 Method 4 (Fay-Herriot Estimator or EBLUP)

Using the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor approach (see *e.g.*, Battese, Harter and Fuller 1988, and Pfeffermann and Barnard 1991), the bias of the synthetic estimator can be reduced considerably by using a composite estimator; for an early reference on composite estimation see Schaible (1978). The composite estimator is obtained as a convex combination of $\underset{\sim}{g}_2$ and a modified $\underset{\sim}{g}_3$. For this purpose, it is assumed that

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta} + \underset{\sim}{a}, \tag{2.4}$$

where a_k's are uncorrelated random small area effects with mean 0 and variance w_k known up to a constant. In our empirical study later we take $w_k = w$. Thus we model $\underset{\sim}{g}_2$ as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{a} + \underset{\sim}{\epsilon}. \tag{2.5}$$

Here $\underset{\sim}{a}$ is also assumed to be uncorrelated with $\underset{\sim}{\epsilon}$. The BLUP of $\underset{\sim}{\Theta}$ under the model defined by (2.4) and (2.5) is

$$\begin{aligned} \underset{\sim}{g}_4 &= \underset{\sim}{g}_3^* + \Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*) \\ &= \Lambda\underset{\sim}{g}_2 + (I - \Lambda)\underset{\sim}{g}_3^*, \end{aligned} \tag{2.6}$$

where

$$\Lambda = (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1},\ U \equiv V + W,$$

$$V = \text{diag}(v_1, \ldots, v_K), \quad W = \text{diag}(w_1, \ldots, w_K),$$

and $\underset{\sim}{g}_3^* = F\underset{\sim}{\beta}^*$, $\underset{\sim}{\beta}^*$ is the WLS estimate of $\underset{\sim}{\beta}$ under model (2.5). Here it is assumed that both the covariance matrices V and W are known in computing the BLUP.

The expression (2.6) follows from the general results on linear models with random effects, see *e.g.*, Rao (1973, p. 267) and Harville (1976). The BLUP or BLUE of $F\underset{\sim}{\beta}$ is $\underset{\sim}{g}_3^*$ and the BLUP of $\underset{\sim}{a}$ is $\Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*)$. It may be of interest to note that the structure of the BLUP does not change regardless of whether or not $\underset{\sim}{\beta}$ is known. However, its MSE does change as expected due to estimation of $\underset{\sim}{\beta}$.

When V and W are replaced by estimates, the estimator $\underset{\sim}{g}_4$ is termed EBLUP. Note that the model (2.4) is more realistic than (2.3), and therefore, the performance of $\underset{\sim}{g}_4$ is expected to be quite favourable. The estimator $\underset{\sim}{g}_4$ approaches $\underset{\sim}{g}_2$ when the v_ks get small, *i.e.*, when the n_{hk}s become large. However, it remains biased, in general, conditional on $\underset{\sim}{\Theta}$, with bias tending to 0 as the v_ks get small.

2.5 Method 5 (Sample Size Dependent Estimator)

An alternative composite estimator is given by the sample size dependent estimator of Drew, Singh and Choudhry (1982). It is defined as

$$\underset{\sim}{g}_5 = \Delta\underset{\sim}{g}_2 + (I - \Delta)\underset{\sim}{g}_3,$$

where $\Delta = \text{diag}(\delta_1, \ldots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{if } \sum_{j \in s_k} d_j \geq \lambda N_k, \\ \sum_{j \in s_k} d_j / \lambda N_k & \text{otherwise} \end{cases} \tag{2.7}$$

and the parameter λ is chosen subjectively as a way of controlling the contribution of the synthetic component. The above estimator takes account of the realized sample size n_{hk}'s and if these are deemed to be sufficiently large according to the condition in (2.7), then it does not rely on the synthetic estimator. This property is somewhat similar to that of $\underset{\sim}{g}_4$; however, unlike $\underset{\sim}{g}_4$, the above estimator does not take account of the relative sizes of the within area and between area variation. Rao and Choudhry (1993) have demonstrated empirically how EBLUPs can sometimes outperform sample size dependent estimators, especially when the between area variation is not large relative to the within area variation. Särndal and Hidiroglou (1989) also proposed estimators similar to the above sample size dependent estimator.

or in matrix notation

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta}, \tag{2.3b}$$

where $F = (F_1, F_2, \ldots, F_K)', F_k = (1, X_k)'$. Now consider a model for the direct small area estimators g_{2k}'s as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

where $\underset{\sim}{g}_2 = (g_{21}, \ldots, g_{2K})', \underset{\sim}{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)', \epsilon_k$s are uncorrelated survey errors with mean 0 and variance v_k. Note that the g_{2k}s are uncorrelated over areas since they are conditionally (given n_{hk}) unbiased and the samples in different small areas are conditionally independent.

Denoting by $\hat{\underset{\sim}{\beta}}$ the weighted least squares (WLS) estimate of $\underset{\sim}{\beta}$, we obtain the regression-synthetic estimator of Θ_k under the assumed model as

$$\underset{\sim}{g}_3 = F\hat{\underset{\sim}{\beta}}.$$

The above estimator could be heavily biased unless the model (2.3) is satisfied reasonably well. The above model may not be realistic because no random fluctuation or random small area effect (a_k, say) is allowed.

2.4 Method 4 (Fay-Herriot Estimator or EBLUP)

Using the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor approach (see *e.g.*, Battese, Harter and Fuller 1988, and Pfeffermann and Barnard 1991), the bias of the synthetic estimator can be reduced considerably by using a composite estimator; for an early reference on composite estimation see Schaible (1978). The composite estimator is obtained as a convex combination of $\underset{\sim}{g}_2$ and a modified $\underset{\sim}{g}_3$. For this purpose, it is assumed that

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta} + \underset{\sim}{a}, \tag{2.4}$$

where a_k's are uncorrelated random small area effects with mean 0 and variance w_k known up to a constant. In our empirical study later we take $w_k = w$. Thus we model $\underset{\sim}{g}_2$ as

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{a} + \underset{\sim}{\epsilon}. \tag{2.5}$$

Here $\underset{\sim}{a}$ is also assumed to be uncorrelated with $\underset{\sim}{\epsilon}$. The BLUP of $\underset{\sim}{\Theta}$ under the model defined by (2.4) and (2.5) is

$$\begin{aligned} \underset{\sim}{g}_4 &= \underset{\sim}{g}_3^* + \Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*) \\ &= \Lambda\underset{\sim}{g}_2 + (I - \Lambda)\underset{\sim}{g}_3^*, \end{aligned} \tag{2.6}$$

where

$$\Lambda = (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1}, U \equiv V + W,$$

$$V = \text{diag}(v_1, \ldots, v_K), \quad W = \text{diag}(w_1, \ldots, w_K),$$

and $\underset{\sim}{g}_3^* = F\underset{\sim}{\beta}^*$, $\underset{\sim}{\beta}^*$ is the WLS estimate of $\underset{\sim}{\beta}$ under model (2.5). Here it is assumed that both the covariance matrices V and W are known in computing the BLUP.

The expression (2.6) follows from the general results on linear models with random effects, see *e.g.*, Rao (1973, p. 267) and Harville (1976). The BLUP or BLUE of $F\underset{\sim}{\beta}$ is $\underset{\sim}{g}_3^*$ and the BLUP of $\underset{\sim}{a}$ is $\Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*)$. It may be of interest to note that the structure of the BLUP does not change regardless of whether or not $\underset{\sim}{\beta}$ is known. However, its MSE does change as expected due to estimation of $\underset{\sim}{\beta}$.

When V and W are replaced by estimates, the estimator $\underset{\sim}{g}_4$ is termed EBLUP. Note that the model (2.4) is more realistic than (2.3), and therefore, the performance of $\underset{\sim}{g}_4$ is expected to be quite favourable. The estimator $\underset{\sim}{g}_4$ approaches $\underset{\sim}{g}_2$ when the v_ks get small, *i.e.*, when the n_{hk}s become large. However, it remains biased, in general, conditional on $\underset{\sim}{\Theta}$, with bias tending to 0 as the v_ks get small.

2.5 Method 5 (Sample Size Dependent Estimator)

An alternative composite estimator is given by the sample size dependent estimator of Drew, Singh and Choudhry (1982). It is defined as

$$\underset{\sim}{g}_5 = \Delta\underset{\sim}{g}_2 + (I - \Delta)\underset{\sim}{g}_3,$$

where $\Delta = \text{diag}(\delta_1, \ldots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{if } \sum_{j \in s_k} d_j \geq \lambda N_k, \\ \sum_{j \in s_k} d_j / \lambda N_k & \text{otherwise} \end{cases} \tag{2.7}$$

and the parameter λ is chosen subjectively as a way of controlling the contribution of the synthetic component. The above estimator takes account of the realized sample size n_{hk}'s and if these are deemed to be sufficiently large according to the condition in (2.7), then it does not rely on the synthetic estimator. This property is somewhat similar to that of $\underset{\sim}{g}_4$; however, unlike $\underset{\sim}{g}_4$, the above estimator does not take account of the relative sizes of the within area and between area variation. Rao and Choudhry (1993) have demonstrated empirically how EBLUPs can sometimes outperform sample size dependent estimators, especially when the between area variation is not large relative to the within area variation. Särndal and Hidiroglou (1989) also proposed estimators similar to the above sample size dependent estimator.

3. METHODS BASED ON POOLED CROSS-SECTIONAL AND TIME SERIES DATA

Suppose information is available for several time points, $t = 1, \ldots, T$, in the form of direct small area estimators $\underset{\sim}{g}_{2t}$, where $\underset{\sim}{g}_{2t}$ is the vector of estimates g_{2k} in (2.2) based on data from time t, and also the small area population totals for the auxiliary variable. We will now introduce some estimators which generalize the Fay-Herriot estimator $\underset{\sim}{g}_{4T}$ in different ways by taking account of the serial dependence of the direct estimates $\{\underset{\sim}{g}_{2t} : t = 1, \ldots, T\}$. Recall that for the Fay-Herriot estimator, the model for $\underset{\sim}{\Theta}_T$ has two components, namely, the structural component $F_T \underset{\sim}{\beta}_T$ and the area component $\underset{\sim}{a}_T$. The estimator $\underset{\sim}{g}_{4T}$ borrows strength over areas for the current time T and is given by the sum of two components, each being EBLUP (BLUE) for the corresponding random (fixed) effect, *i.e.*,

$$\underset{\sim}{g}_{4T} = F_T \underset{\sim}{\beta}^*_T + \underset{\sim}{a}^*_T. \tag{3.1}$$

Methods based on time series data could, however, borrow strength over time as well. Here we introduce three estimators which are motivated from specific structural models for serial dependence. All three of these estimators are optimal under different special cases of a structural time series model for the direct small area estimates $\{\underset{\sim}{g}_{2t} : t = 1, \ldots, T\}$ specified by the following state space model. Let α_t denote $(\underset{\sim}{\beta}'_t, \underset{\sim}{a}'_t)'$ and H_t denote (F_t, I). Then we have

$$\begin{aligned} \underset{\sim}{g}_{2t} &= \underset{\sim}{\Theta}_t + \underset{\sim}{\epsilon}_t, \\ \underset{\sim}{\Theta}_t &= F_t \underset{\sim}{\beta}_t + \underset{\sim}{a}_t \equiv H_t \underset{\sim}{\alpha}_t \end{aligned} \tag{3.2a}$$

and

$$\underset{\sim}{\alpha}_t = G_t \underset{\sim}{\alpha}_{t-1} + \underset{\sim}{\zeta}_t, \tag{3.2b}$$

where

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \underset{\sim}{\zeta}_t = \begin{pmatrix} \underset{\sim}{\xi}_t \\ \underset{\sim}{\eta}_t \end{pmatrix}, \tag{3.2c}$$

along with the usual assumptions about random errors, *i.e.*, $\underset{\sim}{\epsilon}_t$, $\underset{\sim}{\zeta}_t$ are uncorrelated, $\underset{\sim}{\zeta}_t$ is uncorrelated with $\underset{\sim}{\alpha}_s$ for $s < t$, and that $\underset{\sim}{\epsilon}_t \sim (\underset{\sim}{0}, V_t)$, $\underset{\sim}{\zeta}_t \sim (\underset{\sim}{0}, \Gamma_t)$ where $\Gamma_t = \text{block diag}\{B_t, Q_t\}$. The covariance matrices V_t, B_t, and Q_t are generally diagonal. If $G_t^{(1)} = I$ and $G_t^{(2)} = I$ then $\underset{\sim}{\beta}_t$ and $\underset{\sim}{a}_t$ evolve according to a random walk.

This model is in the general class defined by Pfeffermann and Burck (1991) using structural time series models. The main purpose of their study was to show how accounting for cross-sectional correlations between neighbouring small areas (in addition to serial correlations) and inclusion of certain robustness modifications (to protect against model breakdowns) could improve the performance of time series model based estimators. They also used the maximum likelihood method under normality to estimate model parameters. The focus of this paper, on the other hand, is on the Monte Carlo evaluation of a special class of time series estimators (related to Fay-Herriot) chosen on the basis of heuristic considerations and not on the basis of model fitting. The methods considered could, therefore, be viewed as model assisted methods whose performance will be evaluated in a design based (*i.e.*, repeated sampling) framework by Monte Carlo simulation. Moreover, it will be seen later that, for the types of serial dependence considered, the model parameters can be estimated relatively simply by the method of moments, without making any distributional assumptions such as normality.

To find the optimal estimator (BLUP) of $\underset{\sim}{\Theta}_T$ in (3.2) based on all the direct estimates up to time T, we first found the BLUP $\tilde{\underset{\sim}{\alpha}}_T$ of $\underset{\sim}{\alpha}_T$ from which the BLUP of $\underset{\sim}{\Theta}_T$ is obtained as $H_T \tilde{\underset{\sim}{\alpha}}_T$. It is possible, albeit cumbersome, to get $\tilde{\underset{\sim}{\alpha}}_T$ directly from the complete data using the theory of linear models with random effects. However, since the $\underset{\sim}{\alpha}_T$s are connected over time according to the transition equation (3.2b), it is more convenient to compute it recursively using the Kalman filter (KF). Traditionally KF is viewed as a Bayesian technique in which at each time t, the posterior distribution of $\underset{\sim}{\alpha}_t$ given data up to $t - 1$ is updated to get the posterior distribution of $\underset{\sim}{\alpha}_t$ given data up to time t. Although it is instructive to view KF in this manner, it is not necessary under mixed linear models. Suppose $\tilde{\underset{\sim}{\alpha}}_{T|s}$ denotes the BLUP of $\underset{\sim}{\alpha}_T$ based on data up to time s, $s < T$. It is known (see Duncan and Horn 1972) that, for the special structure of serial dependence considered here, the BLUP $\tilde{\underset{\sim}{\alpha}}_T$ of $\underset{\sim}{\alpha}_T$ based on data up to time T is the same as the BLUP of $\underset{\sim}{\alpha}_T$ based on $\tilde{\underset{\sim}{\alpha}}_{T|s}$ and the last $T - s$ observations. In other words, information in the previous data can be condensed into an appropriate BLUP before augmenting more current data points. A good description of the Kalman filter is given in chapter 3 of Harvey (1989).

3.1 Method 6 (Time Series EBLUP-I)

For the first estimator, we let $\underset{\sim}{\beta}_t$ evolve over time (*e.g.*, according to a random walk), but assume that $\underset{\sim}{a}_t$ is serially independent. The equations for the state space model for this case are similar to (3.2) except that the serial independence of the $\underset{\sim}{a}_t$s implies $G_t^{(2)} = 0$. This will give rise to a composite estimator

$$\underset{\sim}{g}_{6T} = F_T \tilde{\underset{\sim}{\beta}}_T + \tilde{\underset{\sim}{a}}_T. \tag{3.3}$$

Note that $\tilde{\underset{\sim}{\beta}}_T$ in (3.3) would now be based on all the small area estimates up to time T and therefore would be different from $\underset{\sim}{\beta}^*_T$ of (3.1) which is based on only direct estimates at time T. The estimator $\tilde{\underset{\sim}{a}}_T$, as a result, would also be different from the corresponding component $\underset{\sim}{a}^*_T$ of (3.1).

In the simulation study described later we take $G_t^{(1)} = I$, $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$, corresponding to a random walk model, and $Q_t = \tau^2 I$. Appendix A.2 illustrates the method of moments estimation of the parameters γ_1^2, γ_2^2, and τ^2. The KF may then be run, with initial values for $\tilde{\underline{\alpha}}_1$ and its MSE obtained from the *FH* estimator at $t = 1$, to obtain the EBLUP of $\tilde{\underline{\alpha}}_T$. Then $H_T \tilde{\underline{\alpha}}_T$ is the time series EBLUP-I estimator $\underline{g}_{6T}$ at time T.

As pointed out by a referee, when the number of small areas is quite large, or when the variation in $\underline{\beta}_t$ over t is relatively large, there is little difference between $\underline{g}_{6T}$ and $\underline{g}_{4T}$. Indeed, there is little difference between the performances of these two estimators in our simulation study described in Section 4.

3.2 Method 7 (Time Series EBLUP-II)

For the second estimator, we let $\underline{\beta}_t$ be fixed (it may or may not be common for different time points) and let the area effects $\underline{a}_t$ be serially dependent according to, for example, a random walk. This time series generalization could be viewed as an analogue of the model proposed by Rao and Yu (1992). The resulting composite estimator will have the same form as (3.1), *i.e.*,

$$\underline{g}_{7T} = F_T \tilde{\underline{\beta}}_T + \tilde{\underline{a}}_T, \tag{3.4}$$

but the component estimates $\tilde{\underline{\beta}}_T$ and $\tilde{\underline{a}}_T$ would be different. We have two cases.

3.2.1 Case 1: Suppose the $\underline{\beta}_t$s are fixed and time-invariant but the $\underline{a}_t$s are serially dependent. Then, in (3.2), $G_t^{(1)} = I$ and $B_t = 0$. If Q_t is taken as $\tau^2 I$, then the only unknown parameter τ^2 can be estimated by the method of moments; see Appendix A.2. We will denote by $\underline{g}_{7T}$ the EBLUP obtained in this case when the parameter estimate is substituted.

3.2.2 Case 2: Here we assume that $\underline{\beta}_t$s are fixed but different for different time points. The area effects $\underline{a}_t$ evolve over time as in Case 1. In (3.2) we have $G_t^{(1)} = 0$ and $B_t = mI$ where m is a large number. The expressions for $\tilde{\underline{\alpha}}_T$ and its MSE obtained from the KF in this case give the correct formulas as $m \to \infty$ (see Sallas and Harville 1981). The KF updating equations for $\tilde{\underline{a}}_t$ in this case take the special form

$$\tilde{\underline{\beta}}_t = (F_t' A_t^{-1} F_t)^{-1} F_t' A_t^{-1} (\underline{g}_{2t} - G_t^{(2)} \tilde{\underline{a}}_{t-1});$$

$$\tilde{\underline{a}}_t = G_t^{(2)} \tilde{\underline{a}}_{t-1} + P_{t|t-1} A_t^{-1} (\underline{g}_{2t} - G_t^{(2)} \tilde{\underline{a}}_{t-1} - F_t \tilde{\underline{\beta}}_t);$$

$$P_t = P_{t|t-1} - P_{t|t-1} A_t^{-1} (A_t - F_t (F_t' A_t^{-1} F_t)^{-1} F_t') A_t^{-1} P_{t|t-1},$$

where $A_t = P_{t|t-1} + V_t$, P_t is the MSE of $\tilde{\underline{a}}_t$ about $\underline{a}_t$, and $P_{t|t-1} = G_t^{(2)} P_{t-1} \{G_t^{(2)}\}' + Q_t$ is the MSE of $G_t^{(2)} \tilde{\underline{a}}_{t-1}$ as an estimator of $\underline{a}_t$. The time series EBLUP in this case will be denoted by $\underline{g}_{7T}^*$.

3.3 Method 8 (Time Series EBLUP-III)

For the third estimator, we let both $\underline{\beta}_t$ and $\underline{a}_t$ evolve over time. This will have more complex serial dependence than either (3.3) or (3.4). Its form will be similar to (3.1) and can be represented as

$$\underline{g}_{8T} = F_T \tilde{\underline{\beta}}_T + \tilde{\underline{a}}_T. \tag{3.5}$$

As before, if $B_t = \text{diag}\{\gamma_1^2, \gamma_2^2\}$ and $Q_t = \tau^2 I$, then the model parameters τ^2, γ_1^2, γ_2^2 can be estimated by the method of moments as in Appendix A.2. The resulting EBLUP of $\underline{\theta}_T$ will be denoted by $\underline{g}_{8T}$.

It may be of interest to note that many of the estimators considered so far are optimal under special cases of the model underlying $\underline{g}_{8T}$. As has been shown, the time series EBLUPs of methods 6 and 7 result from making restrictions on the matrices G_t and Γ_t. The cross-sectional Fay-Herriot estimators of Section 2.4 result from restricting the data to a single time point. The synthetic estimators of section 2.3 are special cases of the Fay-Herriot estimators with zero variance for the random small area effects, and the direct (post-stratified) estimator is obtained in the limit as the variance of the small area effects goes to infinity.

A further generalization that could be useful is to allow correlations between neighbouring small area effects. This can be accomplished by allowing the matrix Q_t in (3.2) to be non-diagonal; however, it is not clear what would be an appropriate correlation structure in Q_t.

4. MONTE CARLO STUDY

The cross-sectional and time series methods were compared empirically by means of a Monte Carlo simulation from a real time series obtained from Statistics Canada's biannual farm surveys, namely, the National Farm Survey (in June) and the January Farm Survey. Due to the redesign after the census of Agriculture in 1986, the survey data for the six time points starting with the summer of 1988 were employed to create a pseudo-population for simulation purposes. To this, data from the census year 1986 was also added. Thus information at one more time point was available although this resulted in a 3-point gap in the series. The missing data points, however, can be easily handled by time series methods. It may be noted that although the data series is short, it is nevertheless believed to be adequate for illustrative purposes. The parameter of interest was taken as the total number of cattle and calves for each crop district (defined as the small area) at each time point. For simplicity, independent stratified random samples were drawn for each occasion from the pseudo-population, though the farm surveys use rotating panels over time. The dependence of direct small area estimates over time was modelled by assuming that the underlying

small area population totals are connected according to some random process. The auxiliary variable used in the model was the ratio-adjusted census 1986 value of the total cattle and calves for each small area. This showed high correlations with the corresponding variable over time at the farm level. Specific details of the empirical study are described below.

4.1 Design of the Simulation Experiment

First we need to construct a pseudo-population from the survey data over six time points (June 1988, January 1989, . . ., January 1991). The actual design involves two frames (list and area) with a one stage stratified sampling from the list frame and a two stage stratified sampling from the area frame, for details see Julien and Maranda (1990). We decided to use survey data from the list frame only because the list frame corresponds to farms existing at the time of Census 1986 and the chosen auxiliary variable for model building was based on Census 1986 information. Moreover, we chose to use the data from the province of Quebec because its area sample is only a minor component of the total sample and the estimated coefficient variation for the twelve crop-districts (*i.e.*, small areas of interest) of this province showed a wide range for the livestock variables. It was decided to avoid variability due to changes in the underlying population over time by retaining only those farms which responded to all the six occasions. Also, farm units who belonged to a multiholding arrangement in any one of the seven time points (including the census) were excluded because of the problems in finding individual farm's data from the multiholding summary record and changes in their reporting arrangement over time.

The various exclusions described above were motivated from considerations of yielding a sharper comparison between small area estimators. The total count of farm units after exclusions was found to be 1,160 out of a total of over 40,000 farms on the list frame. For the pseudo-population, we replicated the 1,160 farm units proportional to their sampling weight so that the total size N of the pseudo-population was 10,362, which was manageable for micro-computer simulation.

The pseudo-population was stratified into four take-some and one take-all strata using Census 1986 count data on cattle and calves as the stratification variable. Although we did not consider alternative stratifications or sample sizes in our simulation study, there is no reason to think that our conclusions would alter significantly if we were to do so. The sigma-gap rule (Julien and Maranda 1990) was used for defining the take-all stratum. To apply the sigma-gap rule we look at the smallest population value greater than the population median where the distance to the next population value, in order of size, is at least one population standard deviation; all units above this point are placed into the take-all stratum. The algorithm of Sethi (1963) was used for determining optimal stratification boundaries for take-some strata. Neyman's optimum allocation was used for sample sizes for strata in order to optimize the precision of the provincial estimate of total count. This resulted in, from a total sample size of 207 (2% sampling rate), allocations of 51, 62, 48 and 35 from takesome strata with 5,001, 3,188, 1,850 and 312 farms, respectively, and the size of the take all stratum was 11. The expected number of sample farms in each small area varied from 4.6 in area 9 up to 27.5 in area 6, with an average of 17.3. The expected number of sample farms with some cattle and calves varied from 3.6 in area 9 to 18.8 in area 3, and the average over the small areas was 11.7. A total of 30,000 simulations were performed. For each simulation, samples were drawn independently for each time point using stratified simple random sampling without replacement. The 30,000 simulations were conducted in 15,000 sets of 2 simulations where each set corresponds to a different vector of realized sample sizes in the twelve small areas within each stratum. This was required to compute certain conditional evaluation measures as described in the next subsection, see also Särndal and Hidiroglou (1989).

4.2 Evaluation Measures

Suppose m simulations are performed in which m_1 sets of different vectors of realized sample sizes in domains (h,k) are replicated m_2 times. The following measures can be used for comparing performance of different estimators at time T. Let i vary from 1 to m_1 and j from 1 to m_2.

(i) Absolute Relative Bias for area k:

$$\mathrm{ARB}_k = \left| m^{-1} \sum_i \sum_j (\mathrm{est}_{ijk} - \mathrm{true}_k)/\mathrm{true}_k \right| . \quad (4.1)$$

The average of ARB_k over areas k will be denoted by AARB. We take the absolute relative bias since our primary interest in this study is in an overall measure like AARB; however, in other contexts the actual biases for individual small areas may also be of considerable interest.

The following measure is motivated by a desire to evaluate the conditional performance of estimators, conditional on the vectors of realized sample sizes in domains. It is conventional to measure performance conditional on fixed domain sample sizes; here we consider the standard deviation of the conditional bias, B_{ik}, as a simple summary measure. If this standard deviation is small then the method is robust to variations in the realized sample sizes. Note that the expected value of B_{ik} is just the unconditional bias which is estimated by ARB_k. Let B_k^2 denote the unconditional expected value of B_{ik}^2. We define the following Monte Carlo measure:

(ii) Standard Deviation of Conditional Relative Bias for area k:

$$\text{SDCRB}_k = \left\{ m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik}) \Big/ \text{true}_k - \text{ARB}_k^2 \right\}^{1/2};$$

$$\hat{B}_{ik} = m_2^{-1} \sum_j \text{est}_{ijk} - \text{true}_k, \qquad (4.2)$$

$$\hat{C}_{ik} = m_2^{-1}(m_2 - 1)^{-1} \left(\sum_j \text{est}_{ijk}^2 - \left(\sum_j \text{est}_{ijk} \right)^2 \Big/ m_2 \right).$$

The correction term $\hat{C}_{ik}$ adjusts for bias in $\hat{B}_{ik}^2$, as an estimate of B_{ik}^2, due to m_2 being finite. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ is conditionally unbiased for B_{ik}^2; it is also unconditionally unbiased for B_k^2. The Monte Carlo average $m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik})$ converges to B_k^2 with probability 1 as $m_1 \to \infty$. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ may be negative for some i, due to finite m_2. For large m_1 the average over i is usually very close to B_k^2; whenever the average is less than ARB_k^2 we set SDCRB_k to 0. ASDCRB will denote the average of SDCRB_k over areas k.

(iii) Mean Absolute Relative Error for area k:

$$\text{MARE}_k = m^{-1} \sum_i \sum_j | \text{est}_{ijk} - \text{true}_k | / \text{true}_k \qquad (4.3)$$

and AMARE denotes the average of MARE_k over areas.

(iv) Mean Squared Error for area k:

$$\text{MSE}_k = m^{-1} \sum_i \sum_j (\text{est}_{ijk} - \text{true}_k)^2 \qquad (4.4)$$

and AMSE as before denotes the average over areas.

(v) Relative Root Mean Squared Error for area k:

$$\text{RRMSE}_k = \{\text{MSE}_k\}^{1/2} / \text{true}_k. \qquad (4.5)$$

Again, ARRMSE denotes the average over areas.

The precision (*i.e.*, the Monte Carlo standard error) of each measure depends on m_1, m_2. For all measures except (ii), the optimal choice of m_1, m_2 under the restriction that $m_2 > 1$ is $m_1 = m/2$, $m_2 = 2$, since this minimizes the Monte Carlo standard error. To see this, let A be the average of an evaluation measure from m_2 samples all with the same sample configuration (set of random sample sizes in domains) which we call C. Then the expected value of A conditional on C is a function of C, say $E(C)$, and the conditional variance of A is proportional to m_2^{-1}, say $V(C)/m_2$. The unconditional variance of A is then $V\{E(C)\} + E\{V(C)\}/m_2$, and the overall Monte Carlo variance of an evaluation measure based on m_1 sample configurations replicated m_2 times is $V\{E(C)\}/m_1 + E\{V(C)\}/m_1 m_2$ which is minimized, since $m = m_1 m_2$ is fixed, by taking m_1 as large as possible. For the second measure, the appropriate choice of m_1, m_2 is less straightforward. In the simulation study, m was chosen as 30,000 and the corresponding values of m_1, m_2 were set at 15,000 and 2.

4.3 Estimators Used in the Comparative Study

There were nine estimators included in the study, namely, g_1 to g_8 and g_7^*, all calculated for time $T = 10$. We used a simple linear regression model for the synthetic component with the auxiliary variable defined as

$$X_{kt} = (\hat{\Theta}_t / \Theta_1) \Theta_{k1}, \qquad (4.6)$$

where Θ_{k1}, Θ_1 respectively denote the population totals for small area k and the province at $t = 1$, *i.e.*, at Census 1986. The estimator $\hat{\Theta}_t$ denotes the post-stratified estimator of Θ_t from the farm survey at time t at the province level. Thus X_{kt} is simply a ratio-adjusted synthetic variable. The variances of error components in the regression model were assumed to be constant over areas. For time series models, it was assumed that the serial dependence was generated by a random walk. The above type of model assumptions have been successfully used in many applications and the main reason for our choice was simplicity. It was hoped, however, that the chosen models might be adequate for our purpose and might illustrate the differential gains with different types of model assisted small area estimators, *i.e.*, both cross-sectional and time series smoothing methods.

Since the Census 1986 data was included in the time series, the direct estimate $\underset{\sim}{g}_{21}$ corresponds to Census 1986 and therefore the survey error $\underset{\sim}{\epsilon}_1$ would be identically $\underset{\sim}{0}$. Moreover, from the definition of X_{kt}, it follows that a reasonable choice of (β_{11}, β_{21}) would be (0,1) which implies that $\underset{\sim}{a}_1$ must be $\underset{\sim}{0}$. Thus the covariance matrices B_t and W_t at $t = 1$ are null and, therefore, the distribution of $\underset{\sim}{\alpha}_t$ at $t = 1$ would not require estimation. The above modification in the initial distribution of $\underset{\sim}{\alpha}_t$ is natural in view of the extra information available from the census. Moreover, since the direct estimates $\underset{\sim}{g}_{2t}$ were not available for $t = 2, 3, 4$, equations for estimating model variance components in Appendix A.2 were modified accordingly.

For method 7 (case 1), $\underset{\sim}{\beta}_t$ was assumed to have a common fixed value only for $t \geq 2$ because at $t = 1$, $\underset{\sim}{\beta}_t = (0,1)'$. For the sample size dependent estimator g_5 the parameter λ was taken to be 1.

4.4 Empirical Results

The main findings were listed in Section 1. Here we give some detailed comparisons and some possible explanations. We do not show separate results for g_7^* which performs slightly worse than, though overall similarly to, g_7. The estimators are summarized in Table 1. Figures 1 to 3 and Tables 2 to 4 present some of the empirical results. We have not shown the Monte Carlo standard errors but they were all found to be quite negligible.

Table 1
Summary of Estimators

g_1 - Expansion	g_6 - Time Series EBLUP-I, βs evolve over time, as independent over time
g_2 - Post-stratified	
g_3 - Synthetic	g_7 - Time Series EBLUP-II, as evolve over time, fixed common β
g_4 - Fay-Herriot	
g_5 - Sample Size Dependent	g_8 - Time Series EBLUP-III, βs and as evolve over time

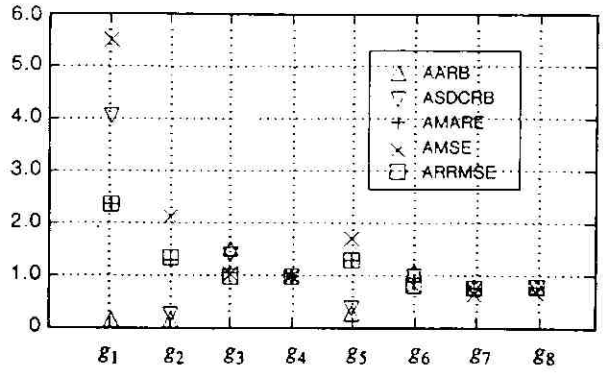

Figure 1. Evaluation Measures Relative to Fay-Herriot
Note: Relative ASDCRB for g_1 (= 18.98) not shown.

Table 2
Average Evaluation Measures

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
AARB	.001	.007	.097	.065	.018	.070	.053	.053
ASDCRB	.282	.016	.016	.015	.023	.010	.010	.010
AMARE	.269	.147	.115	.108	.136	.097	.087	.088
ARRMSE	.339	.192	.137	.137	.176	.120	.109	.111
AMSE (1,000's)	72,979	27,596	13,382	12,898	22,760	10,603	8,610	8,829

Table 2 gives the five evaluation measures averaged over small areas, Figure 1 shows plots of the averaged evaluation measures relative to the Fay-Herriot (g_4) value. There is a clear pattern in the behaviour of various measures across different estimators. The direct estimator g_2 does very well with respect to the bias measure (AARB) but does somewhat poorly with respect to the other measures. The cross-sectional smoothing method g_3 (synthetic) does quite poorly with respect to the bias measures. The Fay-Herriot method g_4 performs somewhat better than post-stratified on average with respect to the MSE measure but is much worse in terms of bias. The sample size dependent method g_5 is quite similar to g_2, slightly worse with respect to the bias measures and slightly better with respect to the other measures. The time series methods g_7 and g_8 perform quite well overall, though they are somewhat worse than g_2 with regard to bias. The performance of the time series estimator g_6 is generally between that of Fay-Herriot and the time series estimators g_7 and g_8. For all of the estimators (including the synthetic g_3) the standard deviation of the conditional relative bias (ASDCRB) is appreciable; however, it is smallest for the time series methods. As expected, the expansion estimator g_1 does well with respect to the unconditional bias measure, AARB, but its conditional performance (ASDCRB) is quite poor.

Figure 2 plots averages of RRMSE_k for three size groups, namely small, medium and large small areas, based on the ranking of their true population totals at time T. They are divided up into these three groups because the relative errors of estimation would be expected to be larger for the smaller totals, and the plots do not contradict this expectation. Again, the time series methods g_7 and g_8 perform best. Note that the time series method g_6, which assumes the small area effects to be independent over time, does not do as well. The unaveraged values of RRMSE_k are given in Table 3. RRMSE_9 is relatively large because the total number of cattle and calves for area 9 is less than half that of any other small area. Areas 6 and 8 stand out within the medium size small areas as being most difficult to estimate by the smoothing methods. The reason for this is that, while there was an overall decline of about 16% in the total number of cattle and calves in the pseudo-population from June 1986 to January 1991, the decreases for areas 6 and 8 were the furthest from the average at 33% and 1%, respectively, so the ratio adjusted covariate would be least appropriate for those areas. Nevertheless, the time series methods g_7 and g_8 performed significantly better than the post-stratified estimator for areas 6 and 8. This is because the random walk model for the small area effects is able to track small areas which, like areas 6 and 8, progressively deviate from the model.

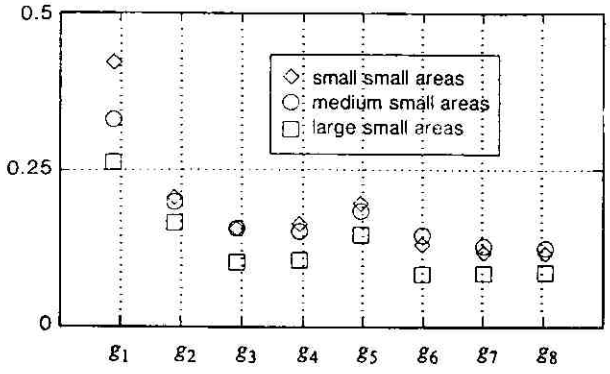

Figure 2. Relative Root Mean Squared Errors: Averaged within Size Groups

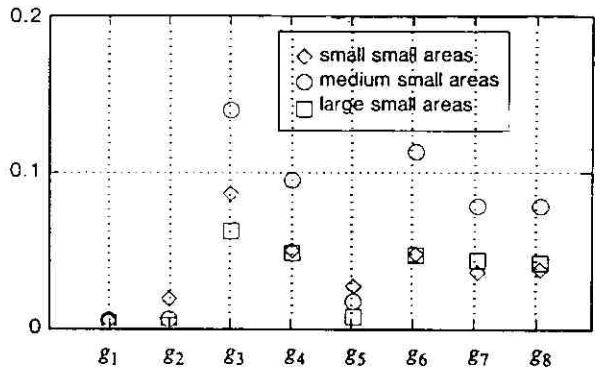

Figure 3. Absolute Relative Biases: Averaged within Size Groups

Table 3

Relative Root Mean Squared Errors and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.580	.277	.342	.275	.277	.199	.160	.174
	10	18,990	.360	.196	.078	.113	.175	.097	.103	.104
	11	18,776	.339	.122	.122	.103	.112	.096	.086	.087
	12	19,819	.409	.237	.076	.152	.212	.123	.117	.117
	Average	16,522	.422	.208	.154	.161	.194	.129	.116	.120
Medium Size	1	27,595	.312	.206	.117	.130	.185	.120	.100	.102
	6	29,012	.306	.241	.256	.216	.224	.224	.168	.172
	7	23,600	.341	.121	.107	.094	.110	.088	.092	.092
	8	23,627	.383	.250	.155	.165	.219	.155	.146	.144
	Average	25,959	.336	.205	.159	.151	.185	.147	.126	.127
Large Size	2	35,592	.268	.171	.113	.110	.156	.096	.089	.088
	3	40,582	.241	.151	.087	.090	.137	.070	.072	.073
	4	42,396	.256	.160	.099	.103	.144	.080	.088	.089
	5	35,996	.270	.176	.091	.097	.160	.088	.085	.088
	Average	38,642	.259	.164	.098	.100	.149	.083	.083	.084

Table 4

Absolute Relative Biases and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.002	.047	.232	.139	.085	.099	.061	.069
	10	18,990	.002	.002	.006	.007	.003	.015	.026	.025
	11	18,776	.002	.009	.090	.052	.021	.062	.039	.037
	12	19,819	.000	.007	.019	.011	.007	.023	.024	.023
	Average	16,522	.001	.016	.087	.052	.029	.050	.037	.039
Medium Size	1	27,595	.001	.003	.093	.063	.007	.078	.044	.045
	6	29,012	.000	.001	.239	.157	.023	.195	.120	.123
	7	23,600	.000	.005	.088	.053	.014	.058	.062	.061
	8	23,627	.002	.008	.143	.106	.024	.124	.093	.091
	Average	25,959	.001	.004	.141	.095	.017	.114	.080	.080
Large Size	2	35,592	.000	.000	.095	.071	.009	.068	.049	.047
	3	40,582	.000	.001	.047	.041	.005	.029	.026	.025
	4	42,396	.001	.002	.066	.056	.008	.044	.057	.056
	5	35,996	.000	.000	.045	.029	.005	.048	.035	.039
	Average	38,642	.000	.001	.063	.049	.006	.047	.042	.042

Figure 3 and Table 4 are identical to Figure 2 and Table 3 in format, but show relative biases instead of relative root mean squared errors. The biases for both the expansion estimator g_1 and the post-stratified g_2 are negligible. For the smoothing methods the average absolute relative biases for medium size small areas are relatively large, mainly because of areas 6 and 8 for which the covariate is least appropriate. Among smoothing methods, the sample size dependent g_5 has the least bias because it is usually very close to the direct g_2; however, it also gains very little over g_2 with respect to mean squared error. Of the remaining smoothing methods the time series estimators g_7 and g_8, which had the smallest mean squared error, also have the smallest bias. Nevertheless, the relative bias of these methods can be quite large, as in areas 6 and 8. In practice it would not be possible to estimate these biases; however, the possible size of the bias could be assessed using simulated sampling from a variety of plausible populations.

5. CONCLUDING REMARKS

It was seen by means of a simulation study that small area estimation methods obtained by combining both cross-sectional and time series data can perform better than those based only on cross-sectional data, with respect to both bias and mean squared error. However, the cost in terms of bias could still be substantial. A question of obvious importance is whether it is possible in practical situations to judge if the gains from any type of smoothing would outweigh the costs, and how to make this judgement.

The models for the simulation study were chosen on general considerations. However, in practice, suitable diagnostics similar to those employed in Pfeffermann and Barnard (1991) should be developed for survey data before any model-assisted method can be recommended. It should also be noted that the small area estimators could be modified to make them robust to mis-specification of the

underlying model as suggested by Pfeffermann and Burck (1990), see also Mantel, Singh and Bureau (1993). Finally, modification and further extension of the methods presented in this paper to the more realistic case of correlated sampling errors should be investigated in the future.

ACKNOWLEDGEMENT

We would like to thank Jon Rao, Danny Pfeffermann and M.P. Singh for useful discussions and comments on earlier versions of this paper. The comments and suggestions of an anonymous referee and an Associate Editor are also very much appreciated. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

APPENDIX

A.1 Variance Estimation for g_{2kt}

Let v_{kt} denote the conditional (given n_{hkt}) variance of g_{2kt} in (2.2). Then v_{kt} is given by (whenever $n_{hkt} > 0$ for all h at time t),

$$v_{kt} = \sum_h N_{hkt}^2 \left(n_{hkt}^{-1} - N_{hkt}^{-1} \right) \sigma_{hkt}^2, \qquad \text{(A.1)}$$

where σ_{hkt}^2 is the population variance for the intersection of the h-th stratum with the k-th small area at time t. The variance σ_{hkt}^2 can be estimated by the usual estimator s_{hkt}^2 for $n_{hkt} \geq 2$. Note that the estimate of the conditional variance v_{kt} also provides an estimate of the unconditional variance of g_{2kt}.

If $n_{hkt} = 1$, then we can use a synthetic value as an estimate of σ_{hkt}^2 which can be defined as $\sum (n_{hkt} - 1) s_{hkt}^2 / \sum (n_{hkt} - 1)$, the summation being over all k for which $n_{hkt} \geq 2$ within each (h,t). If $n_{hkt} = 0$, v_{kt} of (A.1) is of course not defined. With the synthetic value of $\bar{y}_{hkt}$ used in this case, we need a synthetic value of its mean squared error. For each (h,t), it can be defined as

$$(\bar{X}_{hkt}/\bar{X}_{ht})^2 (n_{ht}^{-1} - N_{ht}^{-1}) s_{ht}^2 + (\widehat{\text{bias}})^2,$$

where $(\widehat{\text{bias}})^2$ will be taken as

$$\sum_{n_{hlt}>0} ((\bar{X}_{hlt}/\bar{X}_{ht})\bar{y}_{ht} - \bar{y}_{hlt})^2 / m_{ht},$$

where m_{ht} is the number of small areas with sample in stratum h at time t.

A.2 Estimation of Variance Components

Using the notation of (3.2), we here illustrate the method of moments for estimating variance components for the model of Section 3.1 in the special case when there is only one auxiliary variable X_{ht}, $Q_t = \tau^2 I$ and $\underset{\sim}{\beta}_t$ follows a random walk, *i.e.*, $G_t^{(1)} = I$. Let $F_t = (F_{1t}, \ldots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$, $\underset{\sim}{\beta}_t = (\beta_{1t}, \beta_{2t})'$, and $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$. The parameter τ^2 is estimated by the solution of

$$\sum_{t=1}^{T} \sum_{k=1}^{K} (g_{2kt} - F_{kt}' \hat{\underset{\sim}{\beta}}_t)^2 / (v_{kt} + \tau^2) = T(K - 2).$$

If there is no positive solution, we set $\hat{\tau}^2 = 0$. Here $\hat{\underset{\sim}{\beta}}_t$ denotes the WLS estimate of $\underset{\sim}{\beta}_t$ based on only the cross-sectional data at t. This is analogous to the method used in Fay and Herriot (1979) for cross-sectional data. An estimate of γ_i^2 can be obtained by solving (for $i = 1, 2$)

$$\sum_{t=2}^{T} (\hat{\beta}_{i,t} - \hat{\beta}_{i,t-1})^2 / (\gamma_i^2 + d_{ii}^{(t)}) = T - 1,$$

where $d_{ii}^{(t)}$ is the (i,i)-th element of $(F_{t-1}' U_{t-1}^{-1} F_{t-1})^{-1} + (F_t' U_t^{-1} F_t)^{-1}$.

REFERENCES

BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BELL, W.R., and HILLMER, S.C. (1987). Time series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.

BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

DUNCAN, D.B., and HORN, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.

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

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21, U.S. Office of Management and Budget.

GHOSH, M., and RAO, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, to appear.

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

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: University Press.

JULIEN, C., and MARANDA, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.

MANTEL, H.J., SINGH, A.C., and BUREAU, M. (1993). Benchmarking of small area estimators. *Proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, 920-925.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.

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 Economics Statistics*, 9, 73-84.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. Eds. (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley and Sons.

RAO, J.N.K., and CHOUDHRY, G.H. (1993). Small area estimation: Overview and empirical study. Presented at the International Conference on Establishment Surveys, Buffalo, June 1993, to appear.

RAO, J.N.K., and YU, M. (1992). Small Area Estimation by Combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

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. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

SINGH, A.C., and MANTEL, H.J. (1991) State space composite estimation for small areas. *Proceedings: Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa, November 1991, 17-25.

TILLER, R. (1992). Time series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.

Survey Methodology, December 1994
Vol. 20, No. 2, pp. 171–176
Statistics Canada

Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey

P.D. FALORSI, S. FALORSI and A. RUSSO[1]

ABSTRACT

The study was undertaken to evaluate some alternative small areas estimators to produce level estimates for unplanned domains from the Italian Labour Force Sample Survey. In our study, the small areas are the Health Service Areas, which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut across boundaries of the design strata. We consider the following estimators: post-stratified ratio, synthetic, composite expressed as linear combination of synthetic and of post-stratified ratio, and sample size dependent. For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study in which the sample design was simulated using data from the 1981 Italian Census.

KEY WORDS: Small area estimators; Unplanned domains; Bias; Mean Square Error; Simulation study.

1. INTRODUCTION

In Italy, as in many other countries, there is a growing need for current and reliable data on small areas. This information need concerns most sample surveys realised by the Italian National Statistical Institute (ISTAT), especially the Labour Force Survey (LFS), which has been studied to warrant accuracy in regional estimates.

In the past, ISTAT's solution to this problem was to broaden the sample without changing the estimation method (Fabbris *et al.* 1988). In the last few years, however, in order to find a solution to the negative aspects of oversized samples, research has been launched to identify estimation methods to improve the accuracy of small areas estimates (Falorsi and Russo 1987, 1989, 1990 and 1991).

In our study, the small areas are the Health Service Areas (HSA), which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut-across the boundaries of the design strata. The sizes of these territorial domains are such that the reliability of regular estimates would have been satisfactory had these domains been designed with separate fixed sample sizes from individual domains.

The study was undertaken to evaluate some of the alternative small areas estimators to produce HSA level estimates from the LFS.

We consider the following estimators: post-stratified ratio, synthetic, composite (expressed as linear combination of the synthetic and of the post-stratified ratio), and sample size dependent.

For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study in which the LFS design was simulated using data from the 1981 Italian Census.

2. BRIEF DESCRIPTION OF THE LFS SAMPLE STRATEGY

2.1 Design

The LFS is based on a two stage sample design stratified for the primary sampling units (PSU). The PSUs are the municipalities, while the secondary sampling units (SSU) are the households. In the framework of each geographical region the PSUs are divided according to the provinces. In each province the PSUs are divided into two main area types: the self-representing area consisting of the larger PSUs, and the non self-representing area consisting of the smaller PSUs.

All PSUs in the self-representing area are sampled, while the selection of PSUs in the non self-representing area is carried out within the strata that have approximately equal measures of size. Two sample PSUs are selected from each stratum without replacement and with probability proportional to size (total number of persons). The SSUs are selected without replacement and with equal probabilities from the selected PSUs independently. All members of each sample household are enumerated.

2.2 Estimator of Total

With reference to the generic geographical region, we introduce the following subscripts: h, for stratum ($h = 1, \ldots, H$); i, for primary sampling unit; j, for secondary sampling units; g, for age-sex groups ($g = 1, \ldots, G$).

[1] P.D. Falorsi, Senior Researcher, National Statistical Institute, Rome, Italy; S. Falorsi, Researcher, National Statistical Institute, Rome, Italy; Aldo Russo, Associate Professor, University of Molise, Campobasso, Italy.

In the present study we consider the following age classes 14-19, 20-29, 30-59, 60-64, and over 65.

A quantity referring to stratum h, primary sampling unit i, and secondary sampling unit j will be briefly referred to as the quantity in hij; and a quantity referring to stratum h and primary sampling unit i will be referred to as the quantity in hi.

The following notations are also used: N_h, for number of PSUs in h; P_h, for total number of persons in h; n_h, for number of sample PSUs selected in h; M_{hi} for number of SSUs in hi; P_{hi}, for total number of persons in hi; m_{hi}, for number of sample SSUs selected in hi; P_{ghij}, for number of persons in group g belonging to hij; P_{hij}, for number of persons in hij.

Further let

$$Y = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij}$$

be the total of the characteristic y for regional population, where Y_{ghij} denotes total of the characteristic of interest y for the P_{ghij} persons. Actually, the estimate of Y is obtained by a post-stratified estimator. This estimator is given by:

$$\hat{Y} = \sum_{g=1}^{G} \frac{\hat{Y}_g}{\hat{P}_g} P_g,$$

where

$$\hat{Y}_g = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij}; \; \hat{P}_g = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij}$$

represent unbiased estimates of

$$Y_g = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij} \; ; \quad P_g = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} P_{ghij}.$$

In the above formulas, the symbol K_{hij}, that denotes the basic weight, is expressed by:

$$K_{hij} = \frac{P_h}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}}.$$

3. SMALL AREA ESTIMATORS

With reference to the generic geographical region, we suppose that the population P is divided into D non-overlapping small areas $1, \ldots, d, \ldots, D$ for which estimates are required. Each area is obtained by an aggregation of municipalities. The problem considered is the estimation the total of a y-variable for all units belonging to the small area d. In practice, the small area d will have a non-null intersection with only a certain number of design strata which we denote as $\bar{H} = \{h \mid {}_dP_h > 0\}$, where ${}_dP_h$ represents the part of P_h belonging to the small area d.

Denoting by ${}_dN_h$ the number of PSUs belonging to small area d in stratum h, we seek to estimate the small area total

$${}_dY = \sum_{g=1}^{G} \sum_{h=1}^{\bar{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} Y_{ghij}.$$

The development of a particular estimation method for small areas basically depends on available information. In Italy the accessible information at small area level is very poor. At present the accessible territorial information is total population by sex for each municipality collected through register statistics. In a future context (at end of 1994), the population counts by age-sex group will be available for each municipality. For this reason, in the present study we consider only those small area estimators that utilize, as auxiliary information, the population total by age-sex group.

3.1 Post-stratified Ratio Estimator

A post-stratified ratio estimator (POS) of ${}_dY$ is given by:

$${}_d\hat{Y}_{\mathrm{POS}} = \sum_{g=1}^{G} \frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g} {}_dP_g, \qquad (1)$$

where

$${}_d\hat{Y}_g = \sum_{h=1}^{\bar{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij} \delta_{hi},$$

$${}_d\hat{P}_g = \sum_{h=1}^{\bar{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij} \delta_{hi},$$

$${}_dP_g = \sum_{h=1}^{\bar{H}} {}_dP_{gh} = \sum_{h=1}^{\bar{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} P_{ghij},$$

in which ${}_dP_{gh}$ denotes the total population for the age/sex group g in small area d intersected by stratum h, δ_{hi} is a binary variate that equals 1 if the PSU hi belongs to the small area d and equals 0 otherwise. For a better explanation of formula (1), we observe that PSU is a subset of small area and then does not intersect it.

The post-stratified ratio estimator is unbiased except for the effect of ratio estimation bias which is usually negligible. The estimator is defined to be zero when there is no sample within the domain. This estimator is not reliable for small sample sizes.

3.2 Synthetic Estimator

For computing a synthetic estimator, it is assumed that the small area population means for given population sub-groups are approximately equal to the larger area populations means of the same sub-groups. This estimator is obtained by means of a two steps procedure: (i) with respect to an aggregated territorial level, estimates of the investigated features are determined for population sub-groups; (ii) estimates for the aggregated territorial level area are then scaled in proportion to the sub-group incidence within the small domain of interest.

The synthetic estimator has a low variance since it is based on a larger sample, but it suffers from bias depending on the distance from the assumption of homogeneity, for each subgroup, between the small area and the larger area with reference to the characteristic of interest, y. The problems associated with synthetic estimators have been documented by Purcell and Linacre (1976), Gonzalez and Hoza (1978), Ghangurde and Singh (1978), Schaible (1979) and Levy (1979) among others.

In this study we consider the following form of synthetic estimator (SYN):

$$ {}_d\hat{Y}_{SYN} = \sum_{g=1}^{G} \frac{\tilde{Y}_g}{\tilde{P}_g}\, {}_dP_g, \qquad (2) $$

where

$$ \tilde{Y}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, Y_{ghij};\ \tilde{P}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, P_{ghij}. $$

3.3 Composite Estimator

The composite estimator (COM) considered here is obtained as a linear combination of the estimators SYN (biased with low sample variance) and POS (less biased with high sample variance):

$$ {}_d\hat{Y}_{COM} = \alpha\, {}_d\hat{Y}_{POS} + (1 - \alpha)\, {}_d\hat{Y}_{SYN}, \qquad (3) $$

where α is a constant ($0 \leq \alpha \leq 1$). This estimator minimizes the chances of extreme situations (both in terms of bias and sample variance). Therefore, in a given concrete situation such estimator may turn out to be more advantageous than its two components considered separately.

The optimum value for α that minimizes the MSE of the COM estimator is given by

$$ \alpha_{opt} = \frac{\mathrm{MSE}({}_d\hat{Y}_{SYN}) - E({}_d\hat{Y}_{SYN} - {}_dY)({}_d\hat{Y}_{POS} - Y_d)}{\mathrm{MSE}({}_d\hat{Y}_{SYN}) + \mathrm{MSE}({}_d\hat{Y}_{POS}) - 2E({}_d\hat{Y}_{SYN} - {}_dY)({}_d\hat{Y}_{POS} - Y_d)}. \qquad (4) $$

Furthermore, when neglecting the covariance term in (4), under the assumption that this term will be small relative to $\mathrm{MSE}({}_d\hat{Y}_{SYN})$ and $\mathrm{MSE}({}_d\hat{Y}_{POS})$, the optimal weight α can be approximated by

$$ \alpha^*_{opt} = \frac{\mathrm{MSE}({}_d\hat{Y}_{SYN})}{\mathrm{MSE}({}_d\hat{Y}_{SYN}) + \mathrm{MSE}({}_d\hat{Y}_{POS})}. \qquad (5) $$

This is the approach to define weights followed by Schaible (1978).

In our work the optimal values of α have been obtained from Census data using formula (5). When considering a real sample survey only an estimated value of optimum α may be used, thus resulting in a decrease in efficiency.

3.4 Sample Size Dependent Estimator

The sample size dependent estimator is a particular case of the composite estimator. The linear combination of synthetic and of the less biased estimator is made for each sub-group and depends on the outcome of the given sample. We consider the following form of sample size dependent estimator (SD) which take into account the realized sample size in the small area. It is defined as (Drew, Singh and Choudhry 1982):

$$ {}_d\hat{Y}_{SD} = \sum_{g=1}^{G} \left\{ \alpha_g \left(\frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g}\, {}_dP_g \right) + (1 - \alpha_g) \frac{\tilde{Y}_g}{\tilde{P}_g}\, {}_dP_g \right\}, \qquad (6) $$

where

$$ \alpha_g = \begin{cases} 1/({}_dR_g F) & 1/{}_dR_g < F, \\ 1 & \text{otherwise} \end{cases} \qquad (7) $$

with ${}_dR_g = {}_dP_g / {}_d\tilde{P}_g$.

The constant F is chosen to control the contribution of the synthetic component. The reliance on the synthetic portion decreases as the value of F increases. The choice of the value for F would depend upon several factors. In our study the efficiency of sample dependent estimator has been investigated for $F = 1$. This value proved to be efficient while affording protection against the bias of synthetic estimator.

The logic behind the SD estimator is that when the sample size within domain d and group g is small, then the direct estimate for domain d and group g would be unstable and a synthetic estimate may be superior. However, if the sample in domain d and group g is larger than expected this is not a problem, since the performance of the post-stratified direct part would improve as the sample size improves. In conclusion, we observe that SD estimator may be considered as a particular form of sample size dependent regression estimator given in Särndal and Hidiroglou (1989), that has good conditional properties.

4. DESCRIPTION OF THE EMPIRICAL STUDY

4.1 Simulation of the LFS Sample Design

In our study, we have considered the 14 HSAs of the Friuli region as small areas. The variable of interest, y, is the number of unemployed.

Evaluation of the performance of the various estimators, discussed in Section 3, was done by referring to a sample design (two stages with stratification of the PSUs) identical to that adopted for the LFS in Friuli. This design is based on the selection of 39 PSUs and 2,290 SSUs from a population of 219 PSUs and 465,000 SSUs.

We have selected independently 400 Monte Carlo sample replicates each of identical size (in terms of PSUs and of SSUs) of the LFS' sample. All the information utilized in the simulation is taken from the 1981 General Population Census, so ${}_dY$ is known.

4.2 Evaluation of Small Area Estimators

We denote by ${}_d\hat{Y}(mr)$ the estimate of the total ${}_dY$ for the small area d from the rth Monte Carlo replicate when using the estimator m. The percent relative bias of estimator m for the small area d is given by

$$ {}_d\mathrm{ARB}_m = \frac{1}{R}\left(\sum_{r=1}^{R} \frac{{}_d\hat{Y}(mr)}{{}_dY} - 1 \right) 100, $$

where R is the number of samples ($R = 400$).

The average of the percent absolute relative bias of estimator m over the whole set of small areas is:

$$ \overline{\mathrm{ARB}}_m = \frac{1}{D} \sum_{d=1}^{D} | {}_d\mathrm{ARB}_m |, $$

where D is the number of small areas under observation ($D = 14$).

The percent root mean square error of estimator m for small area d is

$$ {}_d\mathrm{RMSE}_m = \frac{\sqrt{{}_d\mathrm{MSE}_m}}{{}_dY} 100, $$

where the mean square error of estimator m for the small area d is expressed by

$$ {}_d\mathrm{MSE}_m = \frac{1}{R} \sum_{r=1}^{R} ({}_d\hat{Y}(mr) - {}_dY)^2. $$

The average percent root mean square error of estimator m over all areas is

$$ \overline{\mathrm{RMSE}}_m = \frac{1}{D} \sum_{d=1}^{D} {}_d\mathrm{RMSE}_m. $$

4.3 Analysis of Results

A. Overall Performance Measures

The average percent absolute biases and the average percent root mean square errors of the small area estimators for the LFS characteristic "number of unemployed persons" are presented in Table 1. Looking at this table, the following conclusions emerge:

(i) As expected, POS presents the smallest bias. The bias of SYN is larger than the bias of the other estimators. The bias of COM is roughly 30% lower than the bias of SYN estimator. The bias of SD estimator is only slightly lower than that of POS estimator.

(ii) SYN and COM have the smallest average percent root mean square errors, but these estimators are affected by a very high bias. POS, with low bias, is, conversely, the less efficient estimator. The average percent root mean square error of SD is approximately 30% higher than those of SYN and COM estimators.

Table 1

Average Percent Absolute Relative Bias $\overline{\mathrm{ARB}}$ and Average Percent Root Mean Square Error $\overline{\mathrm{RMSE}}$ for Unemployed by Estimator

Estimator	$\overline{\mathrm{ARB}}$	$\overline{\mathrm{RMSE}}$
POS	1.75	42.08
SYN	8.97	23.80
COM	6.00	23.57
SD	2.39	31.08

B. Performance Measures by Small Area

Tables 2 and 3 present the Percent Relative Bias (${}_d$ARB) and the Percent Root Mean Square Error (${}_d$RMSE) of the estimators for each of fourteen Health Service Areas in Friuli. Furthermore, Table 2 gives the percent ratio between the population of the HSA and the population of the set $\tilde{H}$ of strata including the HSA (p_1); Table 3 shows the percent ratio between the population of the HSA and the population of the region Friuli (p_2) and the percent ratio between the population of the set $\tilde{H}$ of strata including the HSA and the population of the region Friuli (p_3). Looking at these Tables, the following conclusions emerge:

(i) SYN and COM are badly biased in some small areas, namely, in those small areas where the model underlying SYN fits poorly. Generally the small areas with low values of the ratio p_1 are affected by large bias (*e.g.*, HSAs 1, 2, 3, 4 and 6). Conversely, large values of the ratio p_1 are associated with low values of the bias (*e.g.*, HSAs 5, 9, 10 and 13). However, SYN and COM consistently have an attractively low RMSE compared to other alternatives. In three of the fourteen areas (*viz*, areas 3, 4 and 8) COM is consistently the most efficient estimator. In two areas (10 and 12)

SYN is evidently more efficient and in the remaining areas the two estimators are roughly similar from the point of view of efficiency. Furthermore, we observe that the lowest values of RMSE for SYN generally are associated with the highest values of the ratio p_3 (*e.g.*, HSAs 1, 2, 5, 6, 9 and 13). HSAs 3 and 4, while having an high value of the ratio p_3, present a high value of RMSE. This is due to the large bias.

(ii) POS shows negligible bias values in almost all small areas. The RMSE values of POS are much higher than those of the other estimators in all the small areas. We observe that the RMSE of the POS estimator is negatively correlated with the ratio p_2. This is caused by the fact that the expected sample size increases as the ratio p_2 increases. Consequently, the variance (which is the main component of MSE of POS) decreases.

(iii) The estimator SD presents a negligible bias in seven (5, 7, 9 10, 11, 12 and 13) of the fourteen small areas. In the other areas the bias is quite low. Furthemore, in nine areas (2, 3, 4, 5, 9, 10, 11, 12 and 13) SD has a bias similar to that of POS. The estimator SD is better, from the MSE point of view, in comparison with POS. In four areas (7, 8, 9, and 13) RMSE is similar to those of SYN and COM.

(iv) Finally, we notice that in the largest areas with the highest values of the ratio p_2 (*e.g.*, HSAs 9 and 5) all the estimators considered give similar results in terms of bias and MSE. For the remaining areas, where the estimators have different performances, there is a problem in the choice of the best estimator.

Table 2

Percent Relative Bias ($_d$ARB) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

		Estimator			
HSA	p_1	POS	SYN	COM	SD
1	19.1	−1.57	−10.92	−7.68	−3.01
2	16.1	−5.61	−9.21	−6.97	−4.79
3	15.3	−5.21	28.82	17.98	5.79
4	16.3	−2.50	20.92	15.02	2.99
5	47.1	−0.46	1.61	0.98	−0.28
6	24.6	−1.37	−12.24	−9.06	−3.28
7	81.8	0.05	−6.25	−3.40	−1.66
8	70.7	0.81	11.80	6.63	2.17
9	92.2	0.47	0.76	0.68	0.78
10	71.2	0.36	−1.34	0.51	−1.02
11	21.7	−1.01	−5.64	−5.00	−1.62
12	40.6	−1.52	−6.66	−6.05	−1.19
13	56.3	−0.95	−3.12	−1.11	−1.28
14	21.8	−2.51	−6.21	−3.03	−3.53

p_1 = percent ratio between the population of the HSA and the population of the set $\bar{H}$ of strata including the HSA.

Table 3

Percent Root Mean Square Error ($_d$RMSE) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

			Estimator			
HSA	p_2	p_3	POS	SYN	COM	SD
1	3.8	19.9	52.23	20.41	21.12	32.39
2	3.1	19.2	63.36	19.45	20.81	38.30
3	3.6	23.2	57.44	36.57	30.71	42.46
4	3.8	23.2	58.19	30.09	27.02	36.88
5	20.2	42.9	18.81	13.38	14.01	17.87
6	8.5	34.8	28.09	17.49	17.00	22.69
7	6.9	8.4	23.83	21.47	21.67	22.67
8	4.8	6.8	28.75	28.54	26.35	27.40
9	21.2	22.9	17.29	16.15	16.40	16.89
10	1.8	2.5	67.00	50.12	53.31	59.27
11	3.2	14.6	49.82	18.35	19.20	30.42
12	4.3	10.7	46.40	22.10	24.04	33.18
13	12.6	22.4	20.13	15.53	15.40	17.88
14	2.3	10.1	57.80	23.58	22.94	36.81

p_2 = percent ratio between the population of the HSA and the population of the region Friuli.

p_3 = percent ratio between the population of the set $\bar{H}$ of strata including the HSA and the population of the region Friuli.

5. CONCLUSIONS

From the point of view of bias, the post-stratified ratio estimator (POS) is essentially unbiased in almost all the small areas. Furthermore the sample size dependent estimator (SD) has negligible values of the bias in almost all small areas. Synthetic (SYN) and composite (COM) estimators present bias values much higher than those of the other estimators.

From the point of view of efficiency, SYN and COM consistently have significantly lower RMSE compared to other alternatives. The estimator SD is much more efficient than POS and furthermore in four of the fourteen areas it shows RMSE values close to those of SYN and COM. Further, when considering the estimator COM there is the problem of the computation of optimum α. In practice only an estimated value of α may be used, resulting in a decrease in efficiency of this estimator. Thus considering both, bias and efficiency, the SD estimator would seem to be preferable to other estimators examined in the context of LFS in Friuli. The sampling rates in Friuli are relatively high and the magnitudes of relative biases and efficiencies of these estimators may be different in other regions where the sampling rates are low, *e.g.*, Piemonte and Lombardia.

REFERENCES

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.

FABBRIS, L., RUSSO, A., and SANETTI, I. (1988). Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 4. Dipartimento di Scienze Statistiche, Universita' di Padova.

FALORSI, P.D., and RUSSO, A. (1987). Un metodo di stima sintetica per piccoli domini territoriali nelle indagini ISTAT sulle famiglie. *Atti del Convegno della Societa' Italiana di Statistica*, Perugia, Italia, 11-20.

FALORSI, P.D., and RUSSO, A. (1989). Un' analisi comparativa di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 18. Dipartimento di Scienze Statistiche, Universita' di Padova.

FALORSI, P.D., and RUSSO, A. (1990). La stima dell'errore quadratico medio di alcune forme di stimatore sintetico nei campioni a due stadi utilizzati nelle indagini ISTAT sulle famiglie. *Giornate di studio: Classificazione ed analisi dei dati, metodi, software, applicazioni:*, Pescara, Italia, 27-39.

FALORSI, P.D., and RUSSO, A. (1991). Evaluation of small area estimation techniques for Italian Labour Force Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 80-106.

GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 52-61.

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.

LEVY, P.S. (1979). Small area estimation synthetic and other procedures, 1968-1978. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 4-19.

PURCELL, N.J., and LINACRE, S. (1976). Techniques for the Estimation of Small Area Characteristics. Paper presented at the 3rd Australian Statistical Conference, Melbourne.

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. (1978). 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. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 36-83.

An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data

GEORGE E. BATTESE, RACHEL M. HARTER, and WAYNE A. FULLER*

Knowledge of the area under different crops is important to the U.S. Department of Agriculture. Sample surveys have been designed to estimate crop areas for large regions, such as crop-reporting districts, individual states, and the United States as a whole. Predicting crop areas for small areas such as counties has generally not been attempted, due to a lack of available data from farm surveys for these areas. The use of satellite data in association with farm-level survey observations has been the subject of considerable research in recent years. This article considers (a) data for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and (b) data obtained from land observatory satellites (LANDSAT) during the 1978 growing season. Emphasis is given to predicting the area under corn and soybeans in these counties. A linear regression model is specified for the relationship between the reported hectares of corn and soybeans within sample segments in the June Enumerative Survey and the corresponding satellite determination for areas under corn and soybeans. A nested-error model defines a correlation structure among reported crop hectares within the counties. Given this model, the mean hectares of the crop per segment in a county is defined as the conditional mean of reported hectares, given the satellite determinations and the realized (random) county effect. The mean hectares of the crop per segment is the sum of a fixed component, involving unknown parameters to be estimated and a random component to be predicted. Variance-component estimators in the nested-error model are defined, and the generalized least-squares estimators of the parameters of the linear model are obtained. Predictors of the mean crop hectares per segment are defined in terms of these estimators. An estimator of the variance of the error in the predictor is constructed, including terms arising from the estimation of the parameters of the model. Predictions of mean hectares of corn and soybeans per segment for the 12 Iowa counties are presented. Standard errors of the predictions are compared with those of competing predictors. The suggested predictor for the county mean crop area per segment has a standard error that is considerably less than that of the traditional survey regression predictor.

KEY WORDS: Small-area estimation; LANDSAT; June Enumerative Survey; Components of variance; Nested-error model.

1. INTRODUCTION

The U.S. Department of Agriculture (USDA) has been investigating the use of LANDSAT satellite data, both to improve its estimates of crop areas for crop-reporting districts and to develop estimates for individual counties. The methodology used in some of these studies was presented by Cárdenas, Blanchard, and Craig (1978), Hanuschak et al. (1979), and Sigman, Hanuschak, Craig, Cook, and Cárdenas (1978). Additional research was presented by Chhikara (1984).

The USDA is engaged in several interrelated types of research. Some research is directed toward transforming satellite information into good estimates of crop areas at the individual pixel and segment levels. The "segment" is the primary sampling unit, and a "pixel" (a term for "picture element") is the unit for which satellite information is recorded. Segments are about 250 hectares; a pixel is about .45 hectares. Other research is aimed at producing good estimators of total crop areas for both large and small geographical units. Studies by Hanuschak et al. (1979) and Hung and Fuller (1987) concentrated on obtaining good functions of the satellite data.

In this article we consider the prediction of areas under corn and soybeans for 12 counties in north-central Iowa, based on 1978 June Enumerative Survey and satellite data. The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in the 37 segments of these 12 counties by interviewing farm operators. Data for more than one sample segment are available for several counties. Based on LANDSAT readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels in the 12 counties. Table 1 presents (a) the number of segments in each county, (b) the number of hectares of corn and soybeans for each sample segment (as reported in the June Enumerative Survey), (c) the number of pixels classified as corn and soybeans for each sample segment, and (d) the county mean number of pixels per segment classified as corn and soybeans.

A preliminary analysis of the corn data indicated that the second segment in Hardin county deviated from other observations: The reported hectares of corn for the second segment were identical to that of the first segment. Therefore, all data for that (second) segment are deleted from our analyses. The soybean data are deleted for convenience, so the same number of observations is involved for both crops.

* George E. Battese is Senior Lecturer, Department of Econometrics, University of New England, Armidale, New South Wales 2351, Australia. Rachel M. Harter is Associate Research Director, A. C. Nielsen Company, Northbrook, IL 60062. Wayne A. Fuller is Distinguished Professor, Department of Statistics, Iowa State University, Ames, IA 50011. This research was partly supported by Research Agreement 58-319T-1-0054X with the Statistical Reporting Service of the U.S. Department of Agriculture, and Joint Statistical Agreement 82-6 with the U.S. Bureau of the Census. The authors thank Cheryl Auer and Stephen Miller for assistance in writing computer programs for the empirical analyses. Comments of the editors and referees resulted in numerous changes to earlier drafts of the article. A part of this research was conducted during the periods the first author was at Iowa State University, on study leaves from the University of New England.

Journal of the American Statistical Association
March 1988, Vol. 83, No. 401, Applications

Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

Figures 1 and 2 plot the reported hectares of corn and soybeans for the remaining 36 segments against the number of pixels of corn and soybeans, respectively. Observations from segments within given counties are identified with different symbols and jointed by lines, so the county data are more clearly indicated. It is evident that there is a strong relationship between the reported hectares of corn and the number of pixels of corn, and between the reported hectares of soybeans and the number of pixels of soybeans. In addition, the plots indicate that observations for segments within counties tend to be closer together than observations for the whole sample.

Predictors of mean crop areas per segment in the sample counties are obtained under the assumption that a linear regression model defines the relationship between the survey and satellite data. The random errors of the model are assumed to be defined by the nested-error model, in which deviations within a county are correlated. Estimation of this model was discussed by Fuller and Battese (1973) and was suggested for small-area estimation by Battese and Fuller (1981) and Fuller and Battese (1981). Alternative approaches to small-area estimation were given by Fuller and Harter (1987). Fuller and Harter (1987) also presented additional details for the methodology in this article.

2. COMPONENTS-OF-VARIANCE MODEL

The reported crop hectares for corn (or soybeans) in sample segments within counties are expressed as a function of the satellite data for those sample segments, such that the reported crop hectares are positively correlated within given counties but uncorrelated from different counties. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij}, \qquad (2.1)$$

where i is the subscript for county ($i = 1, 2, \ldots, T$,

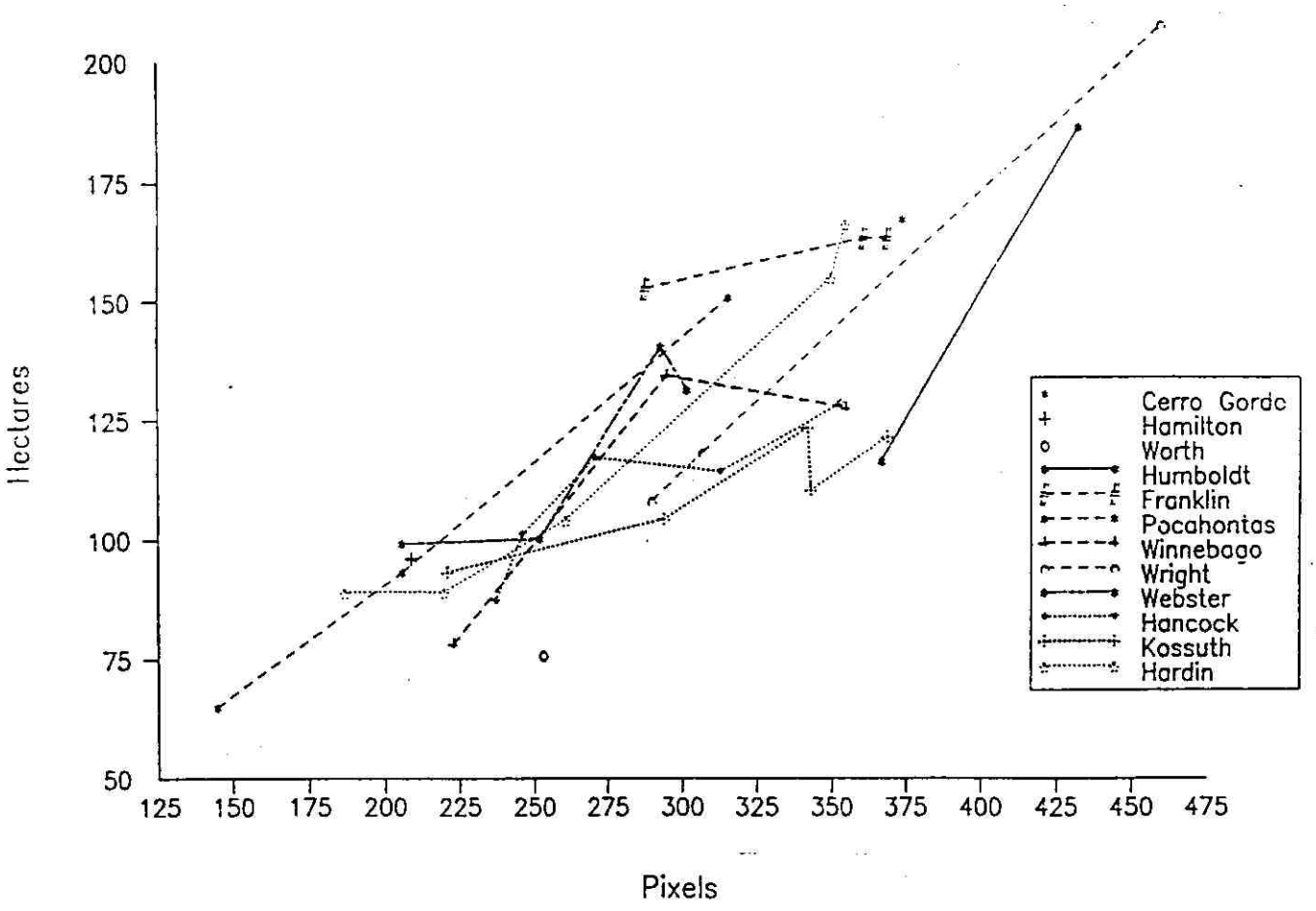

Figure 1. Plot of Corn Hectares Versus Corn Pixels by County.

where $T = 12$); j is the subscript for a segment within a given county ($j = 1, 2, \ldots, n_i$, where n_i is the number of sample segments in the ith county); y_{ij} is the number of hectares of corn (or soybeans) in the jth segment of the ith county, as reported in the June Enumerative Survey; x_{1ij} and x_{2ij} are the number of pixels classified as corn and soybeans, respectively, in the jth segment of the ith county; and β_0, β_1, and β_2 are unknown parameters.

The random error u_{ij}, associated with the reported crop area y_{ij}, is expressed as

$$u_{ij} = v_i + e_{ij}, \tag{2.2}$$

where v_i is the ith county effect and e_{ij} is the random effect associated with the jth sample segment within the ith county. The random errors, v_i ($i = 1, 2, \ldots, T$), are assumed to be iid $N(0, \sigma_v^2)$ random variables independent of the random errors, e_{ij} ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, T$), which are assumed to be iid $N(0, \sigma_e^2)$ random variables. These assumptions imply that the covariance structure of the random errors, u_{ij}, is given by

$$\begin{aligned} E(u_{ij}u_{pq}) &= \sigma_v^2 + \sigma_e^2, && i = p, j = q, \\ &= \sigma_v^2, && i = p, j \neq q, \\ &= 0, && i \neq p. \end{aligned} \tag{2.3}$$

This components-of-variance model is only one possible model for area effects associated with observations from similar geographic regions. Other correlation structures, where reported crop hectares for geographically closer segments have stronger correlation than those farther apart, were considered. Models were estimated where correlation was a function of distance between segments, but the distance effect was not statistically significant.

The components-of-variance model (2.1)–(2.2) does not explicitly define a correlation structure between reported hectares of corn and soybeans in sample segments within counties. The model can be expressed in a multivariate framework that considers the correlation between reported areas of corn and soybeans. Fuller and Harter (1987) covered the multivariate extension of the model (2.1)–(2.3). The extension did not improve the precision of estimation for our data, however, so we confine our attention to the univariate case.

The model for reported hectares of corn (or soybeans), defined by (2.1), was chosen after some preliminary investigations in which the reported hectares of corn (or soybeans) were defined in terms of quadratic functions of the numbers of pixels of corn and soybeans. In each case, however, the null hypothesis—that the coefficients of the nonlinear terms are 0—was not rejected at the 5% level. (Additional evaluation of the model is described in the discussion of empirical results.)

The sample mean of the reported hectares of corn (or soybeans) per segment in the ith county is denoted by

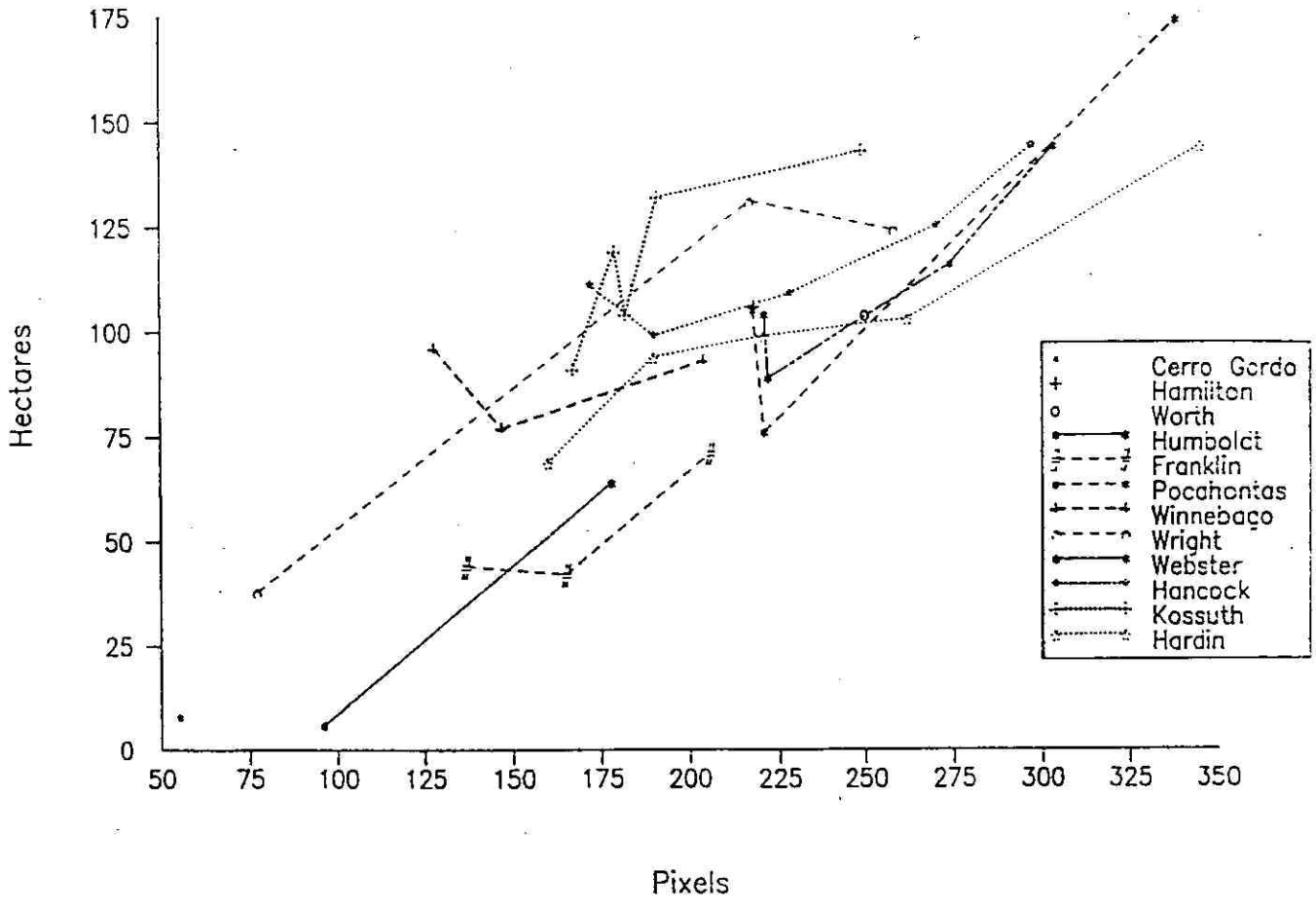

Figure 2. Plot of Soybean Hectares Versus Soybean Pixels by County.

$\bar{y}_{i.}$, where $\bar{y}_{i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$. The sample mean is expressed in terms of the parameters of the model (2.1)–(2.2) by

$$\bar{y}_{i.} = \beta_0 + \beta_1 \bar{x}_{1i.} + \beta_2 \bar{x}_{2i.} + v_i + \bar{e}_{i.}, \qquad (2.4)$$

where $\bar{x}_{1i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{1ij}$ and $\bar{x}_{2i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{2ij}$ are the sample mean numbers of pixels of corn and soybeans, respectively, in the n_i sample segments within county i, and $\bar{e}_{i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ is the sample mean of the within-county effects for the sample segments in the ith county.

The population mean hectares of corn (or soybeans) per segment in the ith county is defined as the conditional mean of the hectares of corn (or soybeans) per segment, given the realized county effect v_i and the values of the satellite data. Under the assumptions of the model (2.1)–(2.2) this mean, denoted by y_i, is

$$y_i \equiv \beta_0 + \beta_1 \bar{x}_{1i(p)} + \beta_2 \bar{x}_{2i(p)} + v_i, \qquad (2.5)$$

where $\bar{x}_{1i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{1ij}$ and $\bar{x}_{2i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{2ij}$ are the population mean numbers of pixels classified as corn and soybeans per segment, respectively, in the ith county, and N_i is the total number of segments in the ith county. Because the number of pixels of corn and soybeans are available from the satellite classifications for all segments in the ith county, the population mean pixel values $\bar{x}_{1i(p)}$ and $\bar{x}_{2i(p)}$ are known. The prediction of the mean crop hectares per segment, defined by (2.5), is the focus of this article.

In a finite-population model, the mean hectares of corn (or soybeans) per segment in the ith county is $\bar{Y}_{i.} \equiv N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$, where Y_{ij} denotes the hectares of the crop in the jth segment in county i and the summation is over all segments in the population. The mean $\bar{Y}_{i.}$ is not equivalent to y_i [as defined in (2.5)], because the sum of the e_{ij}'s over the finite population of segments in county i is not identically 0. As shown in Section 3, however, the predictor for the mean y_i is an appropriate predictor for the finite-population mean $\bar{Y}_{i.}$ when the sampling rate is small.

Obtaining the mean crop hectares per segment in county i, y_i, involves predicting the sum of a known linear function of unknown parameters and an unobserved random variable, v_i. This is a special problem in predicting a linear combination of fixed effects and random effects (see Harville 1976, 1979; Henderson 1975; Kackar and Harville 1984; Peixoto and Harville 1986; Reinsel 1984, 1985). The theory for our parameter estimators and crop-area predictors is an extension of results in the articles cited previously, and is presented in more detail in Fuller and Harter (1987).

Before introducing the estimators and predictors, we present the components-of-variance model (2.1)–(2.3) in matrix notation. Let $\mathbf{Y}_i$ represent the column vector of the reported hectares of the given crop for the n_i sample segments in the ith county, $\mathbf{Y}_i \equiv (y_{i1}, y_{i2}, \ldots, y_{in_i})'$. Furthermore, let $\mathbf{Y}$ represent the column vector of the

reported hectares of the crop for the sample segments in the T counties, $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_T')'$. Thus model (2.1), expressed in matrix notation, is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.6}$$

where the row of $\mathbf{X}$ that corresponds to the element y_{ij} in $\mathbf{Y}$ is $\mathbf{x}_{ij} \equiv (1, x_{1ij}, x_{2ij})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$.

The covariance matrix for the random vector $\mathbf{u}$ in (2.6) is given by

$$E(\mathbf{u}\,\mathbf{u}') \equiv \mathbf{V} = \text{block diag}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_T), \tag{2.7}$$

where

$$\mathbf{V}_i = \mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_\epsilon^2, \tag{2.8}$$

with $\mathbf{J}_i$ the square matrix of order n_i with every element equal to 1 and $\mathbf{I}_i$ the identity matrix of order n_i.

The mean crop hectares per segment (2.5), expressed in matrix notation, is

$$y_i \equiv \bar{\mathbf{x}}_{i(p)}\boldsymbol{\beta} + v_i, \tag{2.9}$$

where $\bar{\mathbf{x}}_{i(p)} \equiv N_i^{-1}\sum_{j=1}^{N_i}\mathbf{x}_{ij} = (1, \bar{x}_{1i(p)}, \bar{x}_{2i(p)})$.

3. ESTIMATION AND PREDICTION

Basic to the prediction of the mean crop area (2.9) for the ith county is the prediction of the county effect, v_i. If the random errors u_{ij} $(j = 1, 2, \ldots, n_i)$ are known, then the best predictor of v_i is the conditional expectation of v_i, given the sample mean $\bar{u}_{i\cdot}$, where $\bar{u}_{i\cdot} \equiv n_i^{-1}\sum_{j=1}^{n_i}u_{ij}$. Under the assumptions of the model (2.1)–(2.2), the random variables v_i and $\bar{u}_{i\cdot}$ have a bivariate normal distribution with zero mean vector and covariance matrix

$$\begin{pmatrix} \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + n_i^{-1}\sigma_\epsilon^2 \end{pmatrix}.$$

The expectation of v_i, conditional on $\bar{u}_{i\cdot}$, is

$$E(v_i|\bar{u}_{i\cdot}) = \bar{u}_{i\cdot}g_i, \tag{3.1}$$

where $g_i = m_i^{-1}\sigma_v^2$ and $m_i = (\sigma_v^2 + n_i^{-1}\sigma_\epsilon^2)$. The error variance in this best predictor is

$$\begin{aligned} E\{(v_i - \bar{u}_{i\cdot}g_i)^2\} &= \sigma_v^2(1 - g_i) \\ &= n_i^{-1}\sigma_\epsilon^2 - n_i^{-2}\sigma_\epsilon^2 m_i^{-1}\sigma_\epsilon^2. \end{aligned} \tag{3.2}$$

If the variances σ_v^2 and σ_ϵ^2 are known, the β parameters of the model can be estimated by the generalized least-squares estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{3.3}$$

Then a possible predictor for the ith county effect, v_i, is

$$\tilde{v}_i = \tilde{u}_{i\cdot}g_i, \tag{3.4}$$

where $\tilde{u}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i}\tilde{u}_{ij}$ and $\tilde{u}_{ij} = y_{ij} - \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}}$. The corresponding predictor $\tilde{y}_i$ for the county mean crop area per segment (2.9) is

$$\tilde{y}_i = \bar{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + \tilde{v}_i. \tag{3.5}$$

This is the best linear unbiased predictor of y_i (see Harville 1985).

The variance of the error in the predictor (3.5) is

$$E\{(\tilde{y}_i - y_i)^2\} = \sigma_v^2(1 - g_i) + \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i', \tag{3.6}$$

where $\mathbf{V}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{c}_i = \bar{\mathbf{x}}_{i(p)} - g_i\bar{\mathbf{x}}_{i\cdot}$, and $\bar{\mathbf{x}}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$. The variance of (3.6) is larger than that of (3.2) by the term associated with the estimation of $\boldsymbol{\beta}$.

A predictor for the finite population mean crop hectares per segment in the ith county [see the paragraph following (2.5)] is

$$N_i^{-1}\left[\sum_{j=1}^{n_i}y_{ij} + \sum_{j=n_i+1}^{N_i}(\mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \tilde{v}_i)\right].$$

In this predictor, the unobserved y_{ij} are replaced by the model predictions. It approaches the predictor (3.5) as the sampling rate decreases. Because the sampling rates are small in our application, we use the predictor (3.5).

This predictor is one of several that have been suggested for the small-area problem. Let a class of predictors of the county mean crop area y_i be defined by

$$\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + (\bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}})\delta_i, \tag{3.7}$$

where δ_i is a nonnegative constant and $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$.

For $\delta_i = g_i$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, the predictor (3.7) is the best linear unbiased predictor. For $\delta_i = 0$, the predictor (3.7) is $\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}}$; this is called the *regression synthetic predictor*. The term *synthetic* is used for predictors that are functions of $\bar{\mathbf{x}}_{i(p)}$, which may not be linear in $\bar{\mathbf{x}}_{i(p)}$. The predictor (3.7) when $\delta_i = 1$ is $\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + (\bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}})$, which is equivalent to the *survey regression predictor* $\bar{y}_{i\cdot} + (\bar{\mathbf{x}}_{i(p)} - \bar{\mathbf{x}}_{i\cdot})\hat{\boldsymbol{\beta}}$. The survey regression predictor adjusts the sample survey mean $\bar{y}_{i\cdot}$, using the difference between the population mean of the regressor vector $\bar{\mathbf{x}}_{i(p)}$ and the sample mean of the regressor values for the sample segments $\bar{\mathbf{x}}_{i\cdot}$ in county i. The survey regression predictor, with an alternative form for the estimator $\hat{\boldsymbol{\beta}}$, was considered by Särndal (1984). Under the model in which $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ the survey regression predictor is unbiased for y_i, conditional on the realized county effect v_i and the values of the satellite data.

The generalized least-squares estimator (3.3) and the predictor (3.5) for the county mean crop area are infeasible, because the variances σ_v^2 and σ_ϵ^2 associated with the nested-error model are unknown. Harville (1977) reviewed a number of methods for estimating the variances for components-of-variance models. We obtain the fitting-of-constants estimator for σ_ϵ^2, denoted by $\hat{\sigma}_\epsilon^2$, which is defined by the residual mean square for the regression model (2.1), with dummy variables for the counties. Alternatively, $\hat{\sigma}_\epsilon^2$ is expressed as

$$\hat{\sigma}_\epsilon^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}\left[\sum_{i=1}^{T}(n_i - 1) - 2\right]^{-1}, \tag{3.8}$$

where $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is the residual sum of squares for the regression of the y deviations, $y_{ij} - \bar{y}_{i\cdot}$, on the x deviations, $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot}$, for those counties with $n_i > 1$. Under the assumptions of model (2.1)–(2.3), the estimator $\hat{\sigma}_\epsilon^2$ is unbiased for σ_ϵ^2 and is distributed as a multiple of a chi-squared random

variable. That is, $d_e\hat{\sigma}_e^2/\sigma_e^2$ has a chi-squared distribution with d_e df, where $d_e \equiv \sum_{i=1}^{T}(n_{i\cdot} - 1) - 2$.

An estimator for the variance of county effects, $\hat{\sigma}_v^2$, is obtained by considering the average of the ordinary least-squares residuals for county i,

$$\hat{\bar{u}}_{i\cdot} = \bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.9}$$

It is readily verified that

$$E(\hat{\bar{u}}_{i\cdot}^2) = b_i\sigma_v^2 + d_i\sigma_e^2, \tag{3.10}$$

where

$$b_i = 1 - 2n_i\bar{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i\cdot}' + \bar{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{j=1}^{T} n_j^2\bar{\mathbf{x}}_{j\cdot}'\bar{\mathbf{x}}_{j\cdot}\right)(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i\cdot}'$$

and $d_i = n_i^{-1}[1 - n_i\bar{\mathbf{x}}_{i\cdot}(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i\cdot}']$. Thus the weighted sum of squares of the average residuals for the counties,

$$\hat{m}_{\cdot\cdot} \equiv \left(\sum_{i=1}^{T} n_i b_i\right)^{-1}\left(\sum_{i=1}^{T} n_i\hat{\bar{u}}_{i\cdot}^2\right), \tag{3.11}$$

has expectation

$$E(\hat{m}_{\cdot\cdot}) \equiv m_{\cdot\cdot} = \sigma_v^2 + c\sigma_e^2, \tag{3.12}$$

where $c = (\sum_{i=1}^{T} n_i b_i)^{-1}(\sum_{i=1}^{T} n_i d_i)$. Under the assumptions of the model (2.1)–(2.2), the weighted sum of squares $\hat{m}_{\cdot\cdot}$ is independent of $\hat{\sigma}_e^2$. Our estimator for σ_v^2 is

$$\hat{\sigma}_v^2 = \max\{\hat{m}_{\cdot\cdot} - c\hat{\sigma}_e^2, 0\}. \tag{3.13}$$

An estimator of g_i is

$$\hat{g}_i = (\hat{\sigma}_v^2 + n_i^{-1}\hat{\sigma}_e^2)^{-1}\hat{\sigma}_v^2. \tag{3.14}$$

A feasible predictor for the mean crop area (2.9) in county i is

$$\hat{y}_i \equiv \bar{\mathbf{x}}_{i(P)}\hat{\boldsymbol{\beta}} + \hat{u}_i\hat{g}_i, \tag{3.15}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated generalized least-squares estimator for $\boldsymbol{\beta}$, obtained by replacing $\mathbf{V}$ of (3.3) with $\hat{\mathbf{V}}$, where $\hat{\mathbf{V}}$ is the estimator for the covariance matrix (2.7) obtained by using the estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, defined by (3.8) and (3.13), respectively; $\hat{u}_{i\cdot} \equiv \bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}}$; and $\hat{g}_i$ is an alternative estimator to (3.14), which is defined in the Appendix. The estimator $\hat{g}_i$, which is approximately unbiased for g_i, was suggested by Fuller and Harter (1987).

An approximation for the variance of the prediction error, $y_i - \hat{y}_i$, and estimators for this variance, were given by Fuller and Harter (1987) for the multivariate case. An estimator for the variance of the prediction error is given in the Appendix. For more detail on the predictor, and the estimator for the variance of the error in the predictor, readers should consult Fuller and Harter (1987).

4. EMPIRICAL RESULTS

Estimates for the parameters of the model (2.1)–(2.2) are obtained by using a modification of the nested-error option of SUPER CARP (see Hidiroglou, Fuller, and Hickman 1980). The modification of SUPER CARP incorporates the alternative estimator for the variance σ_v^2, defined by (3.13). The variance components are first estimated, and then the estimated generalized least-squares estimators for the β parameters are obtained. The estimated parameters for corn are

$$\hat{y}_{ij} = \underset{(25)}{51} + \underset{(.050)}{.329}\,x_{1ij} - \underset{(.056)}{.134}\,x_{2ij},$$

$$\hat{\sigma}_e^2 = \underset{(45)}{150}, \qquad \hat{\sigma}_v^2 = \underset{(89)}{140}.$$

The estimated parameters for soybeans are

$$\hat{y}_{ij} = \underset{(29)}{-16} + \underset{(.058)}{.028}\,x_{1ij} + \underset{(.065)}{.494}\,x_{2ij},$$

$$\hat{\sigma}_e^2 = \underset{(59)}{195}, \qquad \hat{\sigma}_v^2 = \underset{(49)}{272}.$$

The value of the constant c, defined by (3.12), is .349.

The three estimates of the regression function are statistically significant in the corn function, but only the coefficient of soybeans pixels is significantly different from 0 for the soybean function. The estimated variances for within- and among-county variation in reported crop hectares are approximately equal for corn, but for soybeans the among-county variance is about 60% of the total of the two variances. The among-county variance is significant at the 10% level for corn and the 1% level for soybeans.

In our model (2.1)–(2.2) the errors are assumed to be normally distributed. The predictor of the mean crop areas for the counties retains desirable properties for nonnormal errors, but the estimated variances of the prediction errors can be seriously biased when the errors are not normally distributed. Normal probability plots are presented in Figures 3 and 4 for the transformed residuals, $\hat{u}_{ij}^*$, for the corn and soybean models, respectively, which are defined by

$$\hat{u}_{ij}^* = (y_{ij} - \hat{\alpha}_i\bar{y}_{i\cdot}) - (\mathbf{x}_{ij} - \hat{\alpha}_i\bar{\mathbf{x}}_{i\cdot})\hat{\boldsymbol{\beta}},$$

where $\hat{\alpha}_i \equiv 1 - [\hat{\sigma}_e^2/(\hat{\sigma}_e^2 + n_i\hat{\sigma}_v^2)]^{1/2}$. These transformed residuals are approximately uncorrelated with variances approximately equal to σ_e^2 (see e.g., Fuller and Battese 1973, p. 627). The Shapiro–Wilk W statistic for the transformed residuals had values of .985 and .957 for corn and soybeans, respectively. If the residuals were independent normal samples, then the probabilities of values less than those observed would be .921 and .299, respectively. The sample is small, but these analyses give no reason to reject the hypothesis that the errors in the model (2.2) are normally distributed.

Given the assumptions of the model (2.1)–(2.2), the parameters β_1 and β_2 can be estimated from the multiple regression of the within-county deviations $y_{ij} - \bar{y}_{i\cdot}$ on the deviations $x_{1ij} - \bar{x}_{1i\cdot}$ and $x_{2ij} - \bar{x}_{2i\cdot}$ [see Eq. (3.8)]. The expectation of these estimators for β_1 and β_2, represented by $\hat{\boldsymbol{\beta}}_W$, is the same as the expectation of the generalized least-squares estimators of β_1 and β_2, represented by $\hat{\boldsymbol{\beta}}_G$. Hence the estimators $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_G$ can be used to construct a test of model (2.1). Let $\hat{\boldsymbol{\Sigma}}_W$ be the estimated covariance matrix of the within-county estimator $\hat{\boldsymbol{\beta}}_W$, and let $\hat{\boldsymbol{\Sigma}}_G$ be

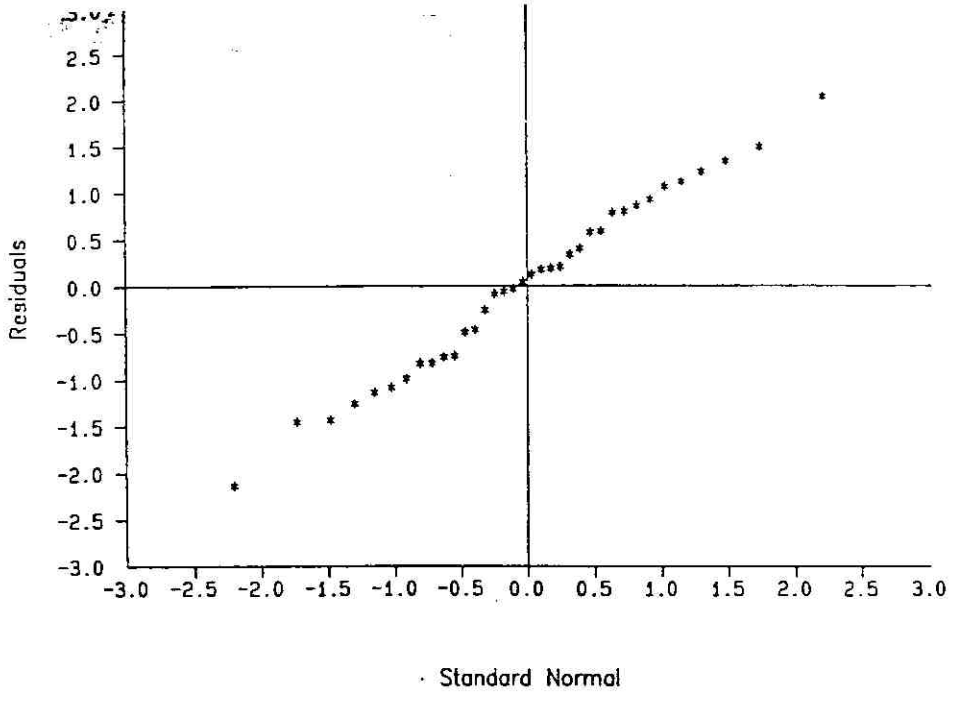

Figure 3. Full Normal Plot Residuals of the Transformed Corn Model.

the estimated covariance matrix of the generalized least-squares estimator $\hat{\beta}_G$. Then the approximate distribution of the statistic

$$F = 2^{-1}(\hat{\beta}_W - \hat{\beta}_G)'(\hat{\Sigma}_W - \hat{\Sigma}_G)^{-1}(\hat{\beta}_W - \hat{\beta}_G)$$

is the F distribution with 2 and 22 df, under the null hypothesis that the slope parameters are the same within and among counties. This result follows from the fact that the estimated covariance between $\hat{\beta}_W$ and $\hat{\beta}_G$ is $\hat{\Sigma}_G$. The test statistic is .46 for corn and .60 for soybeans. Hence we accept the hypothesis that the parameters β_1 and β_2 are the same for within and among counties, as postulated in (2.1).

With the predictor (3.15) we obtain the predictions for the mean crop hectares per segment. Results are given in Tables 2 and 3 for corn and soybeans, respectively, along with the estimated standard errors for the best predictor (3.15), the survey regression predictor, and the sample

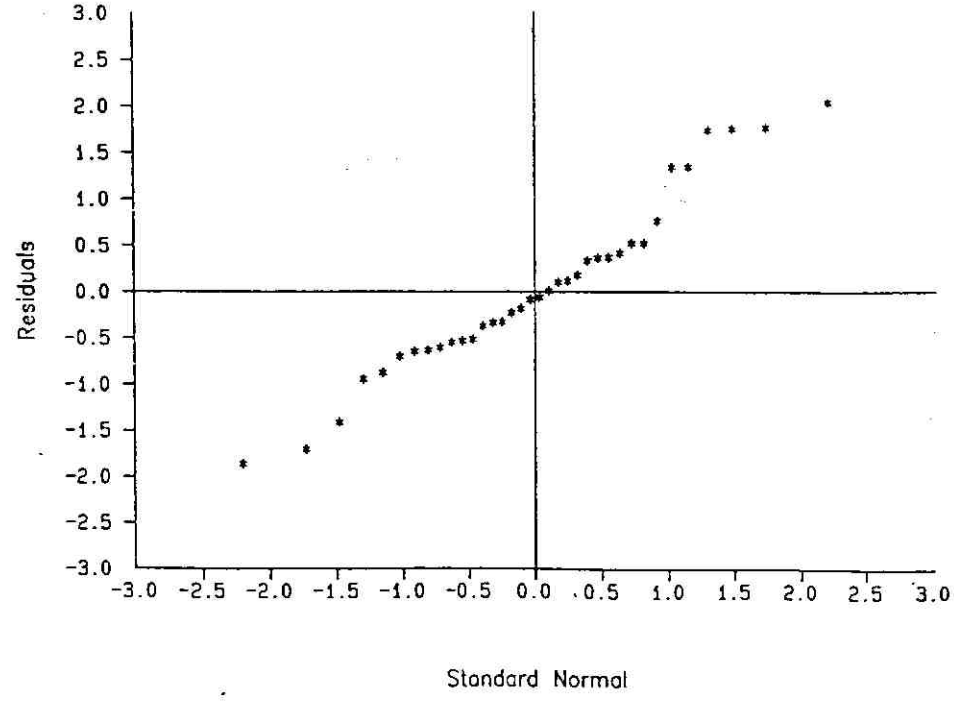

Figure 4. Full Normal Plot Residuals of the Transformed Soybean Model.

Table 2. Predicted Hectares of Corn With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	122.2	9.6	13.7	30.5
Hamilton	1	126.3	9.5	12.9	30.5
Worth	1	106.2	9.3	12.4	30.5
Humboldt	2	108.0	8.1	9.7	21.5
Franklin	3	145.0	6.5	7.1	17.6
Pocahontas	3	112.6	6.6	7.2	17.6
Winnebago	3	112.4	6.6	7.2	17.6
Wright	3	122.1	6.7	7.3	17.6
Webster	4	115.8	5.8	6.1	15.2
Hancock	5	124.3	5.3	5.7	13.6
Kossuth	5	106.3	5.2	5.5	13.6
Hardin	5	143.6	5.7	6.1	13.6

mean of the survey data. The estimated standard error of the sample mean is the square root of the within-county mean square, divided by the number of segments in the given county.

The differences between the predicted hectares of corn and soybeans and the corresponding sample means decrease (see Table 1) as the number of sample segments increases. This is because the standard errors of the sample means are larger for counties with small numbers of sample segments. The standard errors of the sample mean are considerably greater than those for the survey regression predictor. The ratio of the standard error of the best predictor to that for the survey regression predictor increases from about .77 to .97 as the number of sample segments increases from 1 to 5. When there are no more than 3 sample segments, the best predictor has a standard error considerably less than that for the survey regression predictor. The improvement in the precision of the predictor, obtained by increasing the number of sample segments in a county from 3 to 4 or 5, is modest.

5. COMMENTS

The survey regression predictor is unbiased for the 12 counties' mean crop area, and it has relatively small variance. Hence the survey regression predictor is adequate for the entire area. It then becomes desirable to modify the individual county predictors so that the properly weighted sum equals the unbiased survey regression predictor for the total area. A possible adjusted predictor for the ith county mean crop area involves adding to the best predictor a proportion of the weighted sum of the differences between the survey regression predictors and the corresponding best predictor for the counties involved. This predictor is defined by

$$\overset{*}{y}_i = \hat{y}_i + a_i \left[\sum_{j=1}^{T} W_j (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{\boldsymbol{\beta}})(1 - \hat{g}_i) \right],$$

where $a_i = [\sum_{j=1}^{T} W_j^2 \hat{V}(\hat{y}_j)]^{-1} W_i^2 \hat{V}(\hat{y}_i)$. $\hat{V}[\hat{y}_i]$ is the estimated variance of the prediction error for predictor (3.15), and W_j is the weight for the jth area used in constructing the predictor for the total area. It is clear that $\sum_{i=1}^{T} W_i \overset{*}{y}_i$ is equal to the unbiased survey regression predictor for the total area, $\sum_{i=1}^{T} W_i[\bar{y}_{i.} + (\bar{\mathbf{x}}_{(p)} - \bar{\mathbf{x}}_{i.})\hat{\boldsymbol{\beta}}]$. The adjustment produces a very small increase in the variance of the small-area predictors under the components-of-variance model with unequal n_i and/or unequal W_i.

The nested-error regression model (with satellite data as auxiliary variables) offers a promising approach to predicting crop areas in small domains. The USDA has conducted exploratory analyses with the software developed for the univariate nested-error approach to predicting county crop areas. The procedure allows for the use of

Table 3. Predicted Hectares of Soybeans With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	77.8	12.0	15.6	29.1
Hamilton	1	94.8	11.8	14.8	29.1
Worth	1	86.9	11.5	14.2	29.1
Humboldt	2	79.7	9.7	11.1	20.6
Franklin	3	65.2	7.6	8.1	16.8
Pocahontas	3	113.8	7.7	8.2	16.8
Winnebago	3	98.5	7.7	8.3	16.8
Wright	3	112.8	7.8	8.4	16.8
Webster	4	109.6	6.7	7.0	14.6
Hancock	5	101.0	6.2	6.5	13.0
Kossuth	5	119.9	6.1	6.3	13.0
Hardin	5	74.9	6.6	6.9	13.0

supplementary information, such as estimates of variances from other areas and other years, in the estimation of variance components. Modification of the model to account for stratification, according to land use within counties, was investigated by both Walker and Sigman (1982) and Harter (1983).

APPENDIX: COMPUTATIONAL FORMULAS

The model parameters, county predictions, and standard errors in the empirical section were computed with an adaptation of the nested-error regression procedure of SUPER CARP (Hidiroglou, Fuller, and Hickman 1980). The modifications are based on the multivariate estimators suggested by Fuller and Harter (1987). Univariate forms of the estimators for this specific example follow.

The predictor for the county mean crop hectares per segment, defined by (3.15), is

$$\hat{y}_i = \bar{x}_{i(p)}\hat{\beta} + (\bar{y}_i - \bar{x}_i\hat{\beta})\hat{g}_i, \qquad (A.1)$$

where $\hat{g}_i = 1 - \hat{h}_i$,

$$\hat{h}_i = [\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2\hat{w}_i]^{-1}[n_i^{-1}\hat{\sigma}_\epsilon^2 + (n_i^{-1} - c)n_i^{-1}\hat{w}_i],$$

$$\hat{m}_i = \hat{m}_{..} + (n_i^{-1} - c)\hat{\sigma}_\epsilon^2, \qquad \hat{w}_i = 2d_\epsilon^{-1}\hat{m}_i^{-1}\hat{\sigma}_\epsilon^4,$$

$$\hat{k}_i = 2\hat{\sigma}_\epsilon^2(\bar{o}_{ff} + n_i^{-1})^{-1}\left[\sum_{i=1}^{T} n_i b_i\right]^{-2}\left[\sum_{i=1}^{T} n_i^2 b_i(\bar{o}_{ff} + n_i^{-1})^2\right],$$

and $\bar{o}_{ff} = \max[0, (T-5)^{-1}(T-3)\hat{\sigma}_\epsilon^{-2}\hat{m}_{..} - c]$. The constant b_i is defined after (3.10), c is defined after (3.12), and $d_\epsilon = 22$ for this application.

The variance of the error in the predictor (A.1) is estimated by

$$\hat{V}\{\hat{y}_i - y_i\} = n_i^{-1}\hat{\sigma}_\epsilon^2 - \hat{\phi}_i + \hat{c}_i\hat{V}(\hat{\beta})\hat{c}_i' + \hat{h}_i^2\hat{k}_i + d_\epsilon^{-1}\hat{r}_i^2\hat{\phi}_i + d_\epsilon^{-1}\hat{r}_i^2\hat{h}_i\hat{\sigma}_\epsilon^2, \qquad (A.2)$$

where $\hat{c}_i = \bar{x}_{i(p)} - \hat{g}_i\bar{x}_i$, $\hat{\phi}_i = (d_\epsilon + 1)^{-1}d_\epsilon\tilde{\phi}_i - d_\epsilon^{-1}n_i^{-1}\hat{\sigma}_\epsilon^2\hat{h}_i$,

$$\tilde{\phi}_i = n_i^{-2}[\hat{\sigma}_\epsilon^2 + (n_i^{-1} - c)\hat{w}_i]^2[\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2\hat{w}_i]^{-1},$$

and $\hat{r}_i = 1 - (1 - n_i c)\hat{h}_i$. The last three terms of (A.2) are nonnegative and arise from the estimation of the parameter $m_{..}$, defined by (3.12), and the estimation of the variance, σ_ϵ^2.

[*Received June 1984. Revised July 1987.*]

REFERENCES

Battese, G. E., and Fuller, W. A. (1981), "Prediction of County Crop Areas Using Survey and Satellite Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 500–505.

Cárdenas, M., Blanchard, M. M., and Craig, M. E. (1978), *On the Development of Small Area Estimators Using LANDSAT Data as Auxiliary Information*, Washington, DC: Economics, Statistics, and Cooperatives Service, USDA.

Chhikara, R. S. (ed.) (1984), *Communications in Statistics—Theory and Methods*, 13, No. 23 (special issue on crop surveys using satellite data).

Fuller, W. A., and Battese, G. E. (1973), "Transformations of Estimation of Linear Models With Nested-Error Structure," *Journal of the American Statistical Association*, 68, 626–632.

——— (1981), "Regression Estimation for Small Areas," in *Rural America in Passage: Statistics for Policy*, eds. D. M. Gilford, G. L. Nelson, and L. Ingram, Washington, DC: National Academy Press, pp. 572–586.

Fuller, W. A., and Harter, R. M. (1987), "The Multivariate Components of Variance Model for Small Area Estimation," in *Small Area Statistics: An International Symposium*, eds. R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh, New York: John Wiley, pp. 103–123.

Hanuschak, G., Sigman, R., Craig, M., Ozga, M., Luebbe, R., Cook, P., Kleweno, D., and Miller, C. (1979), "Obtaining Timely Crop Area Estimates Using Ground-Gathered and LANDSAT Data," Technical Bulletin 1609, Washington, DC: Economics, Statistics, and Cooperatives Service, USDA.

Harter, R. M. (1983), "Small Area Estimation Using Nested-Error Models and Auxiliary Data," unpublished Ph.D. thesis, Iowa State University, Dept. of Statistics.

Harville, D. A. (1976), "Extension of the Gauss–Markov Theorem to Include the Estimation of Random Effects," *The Annals of Statistics*, 4, 384–395.

——— (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–338.

——— (1979), "Some Useful Representations for Constrained Mixed-Model Estimation," *Journal of the American Statistical Association*, 74, 200–206.

——— (1985), "Decomposition of Prediction Error," *Journal of the American Statistical Association*, 80, 132–138.

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

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP* (6th ed.), Iowa State University, Survey Section, Statistical Laboratory.

Hung, H.-M., and Fuller, W. A. (1987), "Regression Estimation of Crop Acreages With Transformed Landsat Data as Auxiliary Variables," *Journal of Business & Economic Statistics*, 5, 475–482.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–862.

Peixoto, J. L., and Harville, D. A. (1986), "Comparisons of Alternative Predictors Under the Balanced One-Way Random Model," *Journal of the American Statistical Association*, 81, 431–436.

Reinsel, G. (1984), "Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model," *Journal of the American Statistical Association*, 79, 406–414.

——— (1985), "Mean Squared Error Properties of Empirical Bayes Estimators in a Multivariate Random Effects General Linear Model," *Journal of the American Statistical Association*, 80, 642–650.

Särndal, C. E. (1984), "Design-Consistent Versus Model-Dependent Estimation for Small Domains," *Journal of the American Statistical Association*, 79, 624–631.

Sigman, R. S., Hanuschak, G. A., Craig, M. E., Cook, P. W., and Cárdenas, M. (1978), "The Use of Regression Estimation With LANDSAT and Probability Ground Sample Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 165–168.

Walker, G., and Sigman, R. (1982), "The Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston–Ray and the Battese–Fuller Estimators," SRS Staff Report AGES-820909, USDA, Washington, DC.

The Estimation of the Mean Squared Error of Small-Area Estimators

N. G. N. PRASAD and J. N. K. RAO*

Small-area estimation has received considerable attention in recent years because of a growing demand for reliable small-area statistics. The direct-survey estimators, based only on the data from a given small area (or small domain), are likely to yield unacceptably large standard errors because of small sample size in the domain. Therefore, alternative estimators that borrow strength from other related small areas have been proposed in the literature to improve the efficiency. These estimators use models, either implicitly or explicitly, that connect the small areas through supplementary (e.g., census and administrative) data. For example, simple synthetic estimators are based on implicit modeling. In this article, three small-area models, of Battese, Harter, and Fuller (1988), Dempster, Rubin, and Tsutakawa (1981), and Fay and Herriot (1979), are investigated. These models are all special cases of a general mixed linear model involving fixed and random effects, and a small-area mean can be expressed as a linear combination of fixed effects and realized values of random effects. Using the general theory of Henderson (1975) for a mixed linear model, a two-stage estimator (or predictor) of a small-area mean under each model is obtained, by first deriving the best linear unbiased estimator (or predictor) assuming that the variance components that determine the variance–covariance matrix are known, and then replacing the variance components in the estimator with their estimators. Second-order approximation to the mean squared error (MSE) of the two-stage estimator and the estimator of MSE approximation are obtained under normality. Finally, the results of a Monte Carlo study on the efficiency of two-stage estimators and the accuracy of the proposed approximation to MSE and its estimator are summarized. The MSE approximation provides a reliable measure of uncertainty associated with the two-stage estimator. It can also provide asymptotically valid confidence intervals on a small-area mean, as the number of small areas tends to ∞.

KEY WORDS: Best linear unbiased estimator; Fay–Herriot model; Nested error regression model; Random regression coefficient model; Two-stage estimator.

1. INTRODUCTION

In Section 2 we define the three small-area models. The two-stage estimator of a small-area mean under each model is derived in Section 3. A second-order approximation to the mean squared error (MSE) of the two-stage estimator and an estimator of the MSE approximation are obtained under normality in Sections 4 and 5, respectively. The results of a Monte Carlo study on the efficiency of two-stage estimators and the accuracy of the proposed MSE approximation and its estimators are summarized in Section 6.

Although we used Henderson's (1975) approach for a mixed linear model, our results can be restated in the empirical Bayes framework of Morris (1983), Ghosh and Meeden (1986), Ghosh and Lahiri (1987), and others.

2. THREE MODELS

2.1 Nested-Error Regression Model

Battese, Harter, and Fuller (1988) proposed a nested-error regression model in the context of estimating mean acreage under a crop for counties (small areas) in Iowa, using Landsat satellite data in conjunction with survey data. Their model is given by

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \qquad i = 1, \ldots, t, \quad j = 1, \ldots, n_i, \quad (2.1)$$

where y_{ij} is the character of interest for the jth sampled unit in the ith sample area, $\mathbf{x}_{ij} = (x_{ij1} \cdots x_{ijk})'$ is a k vector of corresponding auxiliary values, $\boldsymbol{\beta} = (\beta_1 \cdots \beta_k)'$ is a k vector of unknown parameters, and n_i is the number of sampled units observed in the ith small area ($\Sigma n_i = n$). The random errors v_i are assumed to be independent $N(0, \sigma_v^2)$, independent of the e_{ij}, which are assumed to be independent $N(0, \sigma_e^2)$. The normality assumption is not necessary in deriving the two-stage estimator of a small-area mean. The model (2.1) can be viewed as a random-intercept model by taking $x_{ij1} = 1$ and $\beta_1 = \alpha$. The variables $\alpha_i = \alpha + v_i$ are the random intercepts.

The mean for the ith area may be written as

$$\mu_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + v_i, \quad (2.2)$$

assuming that N_i, the number of population units in the ith area, is large, where $\bar{\mathbf{X}}_i$ is the vector of known means of the $\mathbf{x}_{ij}$ for the ith area. In the application of Battese et al. (1988), N_i ranged from 394 to 965 and n_i from 1 to 6. Note that μ_i is a linear combination of fixed effects $\boldsymbol{\beta}$ and realized value of random effect v_i. It can be interpreted as the conditional mean of y_{ij} for the ith area given v_i.

Finite populations with nonnegligible sampling fractions n_i/N_i are handled by assuming that the N_i units from the ith area are generated from an infinite superpopulation model of the form (2.1); see Sections 3–5.

2.2 Random Regression Coefficient Model

A more general model with random $\boldsymbol{\beta}$ was proposed by Dempster, Rubin, and Tsutakawa (1981). Here, we consider the special case of their model, with single concom-

* N. G. N. Prasad is Assistant Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta T6G 2E1, Canada. J. N. K. Rao is Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank Wayne Fuller, the referees, and the associate editor for valuable comments and many constructive suggestions.

Journal of the American Statistical Association
March 1990, Vol. 85, No. 409, Theory and Methods

itant variable x and regression through the origin. The model may be written as

$$y_{ij} = \beta_i x_{ij} + e_{ij} = \beta x_{ij} + v_i x_{ij} + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, t, \quad (2.3)$$

where $\beta_i = \beta + v_i$ and v_i and e_{ij} are as in the model (2.1). The mean for the ith area is given by

$$\mu_i = \overline{X}_i\beta + \overline{X}_i v_i, \quad (2.4)$$

a linear combination of fixed effect β and realized value of random effect v_i.

2.3 Fay–Herriot Model

In the context of estimating per-capita income for small areas (population less than 1,000), Fay and Herriot (1979) assumed that a k vector of benchmark variables $\mathbf{x}_i = (x_{i1} \cdots x_{ik})'$, related to μ_i, is available for each area i, and that the μ_i are independent $N(\mathbf{x}_i'\boldsymbol{\beta}, A)$, where $\boldsymbol{\beta}$ is a k vector of unknown parameters. They further assumed that the sample mean vector $\bar{\mathbf{y}} = (\bar{y}_1 \cdots \bar{y}_t)' = \mathrm{col}_{1\le i\le t}(\bar{y}_i)$, given $\boldsymbol{\mu} = (\mu_1 \cdots \mu_t)'$, is $N(\boldsymbol{\mu}, \mathbf{D})$, where $\mathbf{D} = \mathrm{diag}(D_1 \cdots D_t)$ with known diagonal elements D_i. The model can be restated as a linear model:

$$\bar{y}_i = \mu_i + e_i \quad \text{and} \quad \mu_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, t, \quad (2.5)$$

where $\mathbf{e} = (e_1 \cdots e_t)'$ and $\mathbf{v} = (v_1 \cdots v_t)'$ are distributed independently as $N(\mathbf{0}, \mathbf{D})$ and $N(\mathbf{0}, A\mathbf{I})$, respectively. The normality assumption is not necessary in deriving the two-stage estimator. Note that the auxiliary information at the unit level is not needed, unlike in the nested-error regression model.

Our results for the Fay–Herriot model are valid for both finite populations with nonnegligible sampling fractions n_i/N_i and general sampling designs. In the latter case, we replace $\bar{y}_i$ by a design-unbiased estimator of $\overline{Y}_i$.

3. TWO-STAGE ESTIMATORS

3.1 Best Linear Unbiased Estimators

The models in Section 2 are all special cases of the general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \quad (3.1)$$

where $\mathbf{y}$ is the vector of sample observations, $\mathbf{X}$ and $\mathbf{Z}$ are known matrices, and $\mathbf{v}$ and $\mathbf{e}$ are distributed independently with means $\mathbf{0}$ and covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively, depending on some parameters $\boldsymbol{\theta}$ called variance components. Henderson (1975) showed that for $\boldsymbol{\theta}$ known the best linear unbiased estimator (or predictor) of $\mu = \mathbf{l}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{v}$ is given by

$$t(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{l}'\tilde{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \quad (3.2)$$

where $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$ is the variance–covariance matrix of $\mathbf{y}$ and $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y})$ is the generalized least squares estimator of $\boldsymbol{\beta}$.

For the nested-error regression model, $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1 \cdots \mathbf{V}_t)$ with $\mathbf{V}_i = \sigma_e^2\mathbf{I}_{n_i} + \sigma_v^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}'$, so $\mathbf{V}^{-1} = \mathrm{diag}(\mathbf{V}_1^{-1} \cdots \mathbf{V}_t^{-1})$ with $V_i^{-1} = (\sigma_e^2)^{-1}\mathbf{I}_{n_i} - \gamma_i n_i^{-1}(\sigma_e^2)^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}'$ and $\gamma_i = \sigma_v^2(\sigma_v^2 + \sigma_e^2 n_i^{-1})^{-1}$, using a standard result on matrix inversion: $(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}/(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u})$ (e.g., see Rao 1973, p. 33). Hence, taking $\mathbf{l} = \overline{\mathbf{X}}_i$ and $\mathbf{m} = (0 \cdots 010 \cdots 0)'$ with 1 in the ith position and noting that $(\sigma_v^2/\sigma_e^2)(1 - \gamma_i) = \gamma_i/n_i$, we get the best linear unbiased estimator of μ_i from (3.2):

$$t_i(\boldsymbol{\sigma}^2, \mathbf{y}) = \overline{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}} + \gamma_i(\bar{y}_i - \bar{\mathbf{x}}_i'\tilde{\boldsymbol{\beta}}), \quad (3.3)$$

where $\boldsymbol{\theta} = \boldsymbol{\sigma}^2 = (\sigma_v^2\ \sigma_e^2)'$ and $\bar{\mathbf{x}}_i$ is the sample mean of $\mathbf{x}_{ij}$ for the ith area. In addition, $\mathbf{y} = \mathrm{col}_{1\le i\le t}\,\mathrm{col}_{1\le j\le n_i}(y_{ij})$ and $\mathbf{X} = \mathrm{col}_{1\le i\le t}\,\mathrm{col}_{1\le j\le n_i}(\mathbf{x}_{ij}')$.

Similar calculations for the random regression coefficient model (2.3) lead to the best linear unbiased estimator of μ_i as

$$t_i(\boldsymbol{\sigma}^2, \mathbf{y}) = \overline{X}_i\tilde{\boldsymbol{\beta}} + \tilde{\gamma}_i\overline{X}_i(\tilde{\beta}_i - \tilde{\beta}). \quad (3.4)$$

Here, $\boldsymbol{\sigma}^2 = (\sigma_v^2\sigma_e^2)'$, $\tilde{\gamma}_i = \sigma_v^2(\sigma_v^2 + \sigma_e^2/\Sigma_j x_{ij}^2)^{-1}$, $\tilde{\beta}_i = \Sigma_j x_{ij}y_{ij}/\Sigma_j x_{ij}^2$, and $\tilde{\beta} = \Sigma\tilde{\gamma}_i\tilde{\beta}_i/\Sigma\tilde{\gamma}_i$.

For the Fay–Herriot model, the best linear unbiased estimator of μ_i is obtained as

$$t_i(A, \bar{\mathbf{y}}) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + (A/(A + D_i))(\bar{y}_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}), \quad (3.5)$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\bar{\mathbf{y}}$, with $\mathbf{V} = \mathrm{diag}(A + D_1 \cdots A + D_t)$ and $\mathbf{X} = \mathrm{col}_{1\le i\le t}(\mathbf{x}_i')$. Under normality, the estimator (3.5) is also a Bayes estimator, as shown by Fay and Herriot (1979). Note that $t_i(A, \bar{\mathbf{y}})$ tends to the direct survey estimator $\bar{y}_i$ as $D_i/(A + D_i) \to 0$ and to the synthetic estimator $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}$ as $A/(A + D_i) \to 0$. Thus the best linear unbiased estimator is a weighted average $w_i\bar{y}_i + (1 - w_i)\mathbf{x}_i'\tilde{\boldsymbol{\beta}}$, where $w_i = A/(A + D_i)$ reflects the uncertainty, A, in the model for the μ_i relative to the total variance $A + D_i$.

Assuming a superpopulation model of the form (2.1) for the N_i population units in the ith area, it can be shown that the best linear unbiased estimator of $\overline{Y}_i = \Sigma_j y_{ij}/N_i$ under the nested-error regression model is given by

$$t_i^F(\boldsymbol{\sigma}^2, \mathbf{y}) = f_i\bar{y}_i + (1 - f_i)t_i^*(\boldsymbol{\sigma}^2, \mathbf{y}), \quad (3.6)$$

where $f_i = n_i/N_i$, $t_i^*(\boldsymbol{\sigma}^2, \mathbf{y})$ is given by (3.3), with $\mathbf{X}_i$ replaced by $\bar{\mathbf{x}}_i^*$, the mean of the $\mathbf{x}_{ij}$ for the $N_i - n_i$ nonsampled units, and F stands for finite populations. Similarly, for the random regression coefficient model the best linear unbiased estimator of $\overline{Y}_i$ is given by (3.6), with $t_i^*(\boldsymbol{\sigma}^2, \mathbf{y})$ changed to (3.4) and $\overline{X}_i$ replaced by $\bar{x}_i^*$.

3.2 Two-Stage Estimators

The estimator $t(\boldsymbol{\theta}, \mathbf{y})$ [written as $t(\boldsymbol{\theta})$ for convenience] depends on the variance components $\boldsymbol{\theta} = (\theta_1 \cdots \theta_p)'$, but in practice the components of $\boldsymbol{\theta}$ will be unknown. Actually, it depends solely on the ratios θ_i/θ_p; for example, (3.3) and (3.4) depend solely on σ_v^2/σ_e^2. It is customary to estimate $t(\boldsymbol{\theta})$ by replacing $\boldsymbol{\theta}$ with an asymptotically consistent estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$. The resulting two-stage estimator, $t(\hat{\boldsymbol{\theta}})$, remains unbiased for μ (i.e., $E[t(\hat{\theta}) - \mu] = 0$), provided that $E[t(\theta)]$ is finite, the elements of $\hat{\boldsymbol{\theta}}$ are even functions of $\mathbf{y}$ and translation-invariant [i.e., $\hat{\boldsymbol{\theta}}(-\mathbf{y}) = \hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y} - \mathbf{X}\mathbf{a}) = \hat{\boldsymbol{\theta}}(\mathbf{y})$ for all $\mathbf{y}$ and $\mathbf{a}$], and the

distributions of $\mathbf{v}$ and $\mathbf{e}$ are both symmetric (not necessarily normal); see Kackar and Harville (1984). Nevertheless, the MSE of $t(\hat{\boldsymbol{\theta}})$ will increase relative to the MSE of $t(\boldsymbol{\theta})$ (see Sec. 4).

Various methods of estimating $\boldsymbol{\theta}$ for a general mixed linear model are available (see Harville 1977 for an excellent review), but here we confine ourselves to the well-known method of fitting constants, called Henderson's method 3.

For the nested-error regression model, unbiased quadratic estimators of σ_e^2 and σ_v^2 from Henderson's method 3 are given by

$$\hat{\sigma}_e^2 = (n - t - k + \lambda)^{-1}\Sigma\Sigma\hat{e}_{ij}^2 \tag{3.7}$$

and

$$\tilde{\sigma}_v^2 = n_*^{-1}[\Sigma\Sigma\hat{u}_{ij}^2 - (n - k)\hat{\sigma}_e^2], \tag{3.8}$$

where $n_* = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^{t} n_i^2\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i']$, and $\lambda = 0$ if the model (2.1) has no intercept term and $\lambda = 1$ otherwise. Furthermore, $\{\hat{e}_{ij}\}$ are the residuals from the ordinary least squares regression of $y_{ij} - \bar{y}_i$ on $\{x_{ij1} - \bar{x}_{i1} \cdots x_{ijk} - \bar{x}_{ik}\}$ and $\{\hat{u}_{ij}\}$ are the residuals from the ordinary least squares regression of y_{ij} on $\{x_{ij1} \cdots x_{ijk}\}$.

For the random regression coefficient model (2.3), unbiased quadratic estimators of σ_e^2 and σ_v^2 are given by

$$\hat{\sigma}_e^2 = (n - t)^{-1}\Sigma\Sigma\hat{e}_{ij}^2 \tag{3.9}$$

and

$$\tilde{\sigma}_v^2 = \bar{n}_*^{-1}[\Sigma\Sigma\hat{u}_{ij}^2 - (n - 1)\hat{\sigma}_e^2], \tag{3.10}$$

with $\hat{e}_{ij} = y_{ij} - x_{ij}(\Sigma_j x_{ij}y_{ij})(\Sigma_j x_{ij}^2)^{-1}$ and $\bar{n}_* = \Sigma\Sigma x_{ij}^2 - [\Sigma_i(\Sigma_j x_{ij}^2)^2](\Sigma\Sigma x_{ij}^2)^{-1}$.

An unbiased quadratic estimator of A in the Fay–Herriot model is given by

$$\hat{A} = (t - k)^{-1}\left[\sum_{i=1}^{t}\hat{u}_i^2 - \sum_{i=1}^{t} D_i(1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i)\right], \tag{3.11}$$

where $\hat{u}_i = \bar{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{y}}$. The two-stage estimator $t_i(\hat{A}, \bar{\mathbf{y}})$ is an empirical Bayes estimator of μ_i under normality (Fay and Herriot 1979).

It is possible for $\tilde{\sigma}_v^2$ [defined by (3.8) and (3.10)] or $\hat{A}$ [given by (3.11)] to take negative values, but $\Pr(\tilde{\sigma}_v^2 \leqq 0)$ or $\Pr(\hat{A} \leqq 0)$ tends to 0 as $t \to \infty$. If $\tilde{\sigma}_v^2$ or $\hat{A}$ is negative, we set it equal to 0, which ensures that the two-stage estimator has a finite expectation. We define $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ and $\hat{A} = \max(\hat{A}, 0)$. Generally, we denote the Henderson unbiased estimator of $\boldsymbol{\theta}_i$ by $\tilde{\boldsymbol{\theta}}_i$ and define $\hat{\theta}_i = \max(\tilde{\theta}_i, 0)$.

4. SECOND-ORDER APPROXIMATION TO MSE

Kackar and Harville (1984) showed that

$$\text{MSE}[t(\hat{\boldsymbol{\theta}})] = \text{MSE}[t(\boldsymbol{\theta})] + E[t(\hat{\boldsymbol{\theta}}) - t(\boldsymbol{\theta})]^2, \tag{4.1}$$

under normality, provided that $\hat{\boldsymbol{\theta}}$ is translation-invariant. Henderson (1975) gave an expression for $\text{MSE}[t(\boldsymbol{\theta})]$, but the second term of (4.1) is generally not tractable except in special cases, such as the balanced one-way analysis of variance model $y_{ij} = \mu + v_i + e_{ij}$, with $n_i = r$ (Peixoto and Harville 1986). Kackar and Harville (1984) obtained a Taylor series approximation

$$E[t(\hat{\boldsymbol{\theta}}) - t(\boldsymbol{\theta})]^2 \doteq E[\mathbf{d}(\boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 \tag{4.2}$$

with $\mathbf{d}(\boldsymbol{\theta}) = \partial t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and then proposed a further approximation

$$E[\mathbf{d}(\boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 \doteq \text{tr}[\mathbf{A}(\boldsymbol{\theta})E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'], \tag{4.3}$$

where $\mathbf{A}(\boldsymbol{\theta})$ is the covariance matrix of $\mathbf{d}(\boldsymbol{\theta})$.

We propose a further approximation, given by

$$\text{tr}[\mathbf{A}(\tilde{\boldsymbol{\theta}})E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'] \doteq \text{tr}[(\nabla\mathbf{b})\mathbf{V}(\nabla\mathbf{b}')E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'], \tag{4.4}$$

where $\nabla\mathbf{b}' = \text{col}_{1\leq j\leq p}(\partial\mathbf{b}'/\partial\theta_j)$ and $\mathbf{b}' = \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$. General conditions are given in the Appendix (Theorem A.1), under which the precise order of neglected terms in the approximations (4.3) and (4.4) is $o(t^{-1})$ for large t. The three small-area models satisfy these conditions. We also show in the Appendix (Theorem A.2) that the precise order of neglected terms in the Taylor series approximation (4.2), under the Fay–Herriot model (2.5), is $o(t^{-1})$. It seems more difficult to give general conditions as in the case of Approximations (4.3) and (4.4), but the proof essentially involves showing that $Eo_p(t^{-1}) = o(t^{-1})$, where $o_p(t^{-1})$ denotes terms of lower order than t^{-1} in probability.

Combining (4.1)–(4.4), we get

$$\text{MSE}[t(\hat{\boldsymbol{\theta}})] \doteq \text{MSE}[t(\boldsymbol{\theta})] + \text{tr}[(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')'E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})']. \tag{4.5}$$

We now evaluate (4.5) for each of the three small-area models. Using Henderson's (1975) general result on $\text{MSE}[t(\boldsymbol{\theta})]$ or by direct calculation, $\text{MSE}[t_i(\boldsymbol{\sigma}^2, \mathbf{y})]$ for the nested-error regression model is obtained as

$$\text{MSE}[t_i(\boldsymbol{\sigma}^2, \mathbf{y})] = (1 - \gamma_i)\sigma_v^2 + (\bar{\mathbf{X}}_i - \gamma_i\bar{\mathbf{x}}_i)'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\bar{\mathbf{X}}_i - \gamma_i\bar{\mathbf{x}}_i), \tag{4.6}$$

under the arbitrary distributions of $\{v_i\}$ and $\{e_{ij}\}$, where $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ is the variance–covariance matrix of $\tilde{\boldsymbol{\beta}}$.

Noting that $(\sigma_v^2/\sigma_e^2)(1 - \gamma_i) = \gamma_i/n_i$ and

$$(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')' = n_i^{-2}(\mathbf{1}_{n_i}'\mathbf{V}_i\mathbf{1}_{n_i})\begin{bmatrix}\partial\gamma_i/\partial\sigma_v^2\\ \partial\gamma_i/\partial\sigma_e^2\end{bmatrix}[\partial\gamma_i/\partial\sigma_v^2\ \delta\gamma_i/\partial\sigma_e^2],$$

where $\mathbf{1}_{n_i}'\mathbf{V}_i\mathbf{1}_{n_i} = n_i^2(\sigma_v^2 + \sigma_e^2/n_i)$, we get

$$\text{tr}[(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')'E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'] = n_i^{-2}(\sigma_v^2 + \sigma_e^2/n_i)^{-3}\text{var}(\tilde{\sigma}_e^2\sigma_v^2 - \tilde{\sigma}_v^2\sigma_e^2). \tag{4.7}$$

Similar calculations under the random regression coefficient model give

$$\text{MSE}[t_i(\boldsymbol{\sigma}^2, \mathbf{y})] = \bar{X}_i^2(1 - \tilde{\gamma}_i)\sigma_v^2 + \bar{X}_i^2(1 - \tilde{\gamma}_i)^2(\Sigma\tilde{\gamma}_i)^{-1}\sigma_v^2 \tag{4.8}$$

and

$$\text{tr}[(\nabla \mathbf{b}')\mathbf{V}(\nabla \mathbf{b}')'E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'] = \bar{X}_i^2(\Sigma_j x_{ij}^2)^{-2}(\sigma_v^2 + \sigma_e^2/\Sigma_j x_{ij}^2)^{-3}\text{var}(\tilde{\sigma}_e^2\sigma_v^2 - \tilde{\sigma}_v^2\sigma_e^2). \quad (4.9)$$

In the case of the Fay–Herriot model, we get

$$\text{MSE}[t_i(A, \bar{\mathbf{y}})] = AD_i(A + D_i)^{-1} + D_i^2(A + D_i)^{-2}\mathbf{x}_i'(X'V^{-1}X)^{-1}\mathbf{x}_i \quad (4.10)$$

and

$$\text{tr}[(\nabla \mathbf{b}')\mathbf{V}(\nabla \mathbf{b}')'E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'] = D_i^2(A + D_i)^{-3}\text{var}(\hat{A}). \quad (4.11)$$

Ignoring the uncertainty in $\hat{\sigma}^2$ or $\hat{A}$ and using the MSE of the best linear unbiased estimator of μ_i as an approximation to the MSE of the corresponding two-stage estimator of μ_i could lead to a serious understatement, since the neglected terms are of the same order, $O(t^{-1})$, as the term obtained by estimating β in the MSE of the best linear unbiased estimator.

For finite populations, the MSE of $t_i^F(\hat{\sigma}^2, \mathbf{y})$ is given by

$$\text{MSE}[t_i^F(\hat{\sigma}^2, \mathbf{y})] = (1 - f_i)^2[\text{MSE}(t_i^*(\hat{\sigma}^2, \mathbf{y})) + N_i^{-1}(1 - f_i)^{-1}\sigma_e^2], \quad (4.12)$$

noting that $t_i^F(\hat{\sigma}^2, \mathbf{y}) - \bar{Y}_i = (1 - f_i)[\{t_i^*(\hat{\sigma}^2, \mathbf{y}) - \mu_i\} - \bar{e}_i^*]$, where $\bar{e}_i^*$ is the mean of the e_{ij} for the $N_i - n_i$ nonsampled units. The approximation to MSE of $t_i^*(\hat{\sigma}^2, \mathbf{y})$ is the sum of (4.6) and (4.7) [or (4.8) and (4.9)], with $\bar{X}_i$ replaced by $\bar{x}_i^*$ in (4.6). Hence no new theory is required for getting a second-order approximation to the MSE of the two-stage estimator $t_i^F(\hat{\sigma}^2, \mathbf{y})$. For the special case of a random one-way model, $y_{ij} = \mu + v_i + e_{ij}$, the estimator $t_i^F(\hat{\sigma}^2, \mathbf{y})$ is essentially identical to the empirical Bayes estimator of $\mu_i = \mu + v_i$ proposed by Ghosh and Meeden (1986) and Ghosh and Lahiri (1987).

5. ESTIMATOR OF MSE APPROXIMATION

We now obtain estimators of the MSE approximation (4.5) for the three models under normality. The Appendix (Theorem A.3) shows that the expectation of the MSE estimator is correct to $O(t^{-1})$ under the Fay–Herriot model. Again, it seems more difficult to give general conditions, but the proof essentially involves showing that $Eo_p(t^{-1}) = o(t^{-1})$.

For the nested-error regression model, the MSE approximation may be written as $g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + g_{3i}(\sigma^2)$, where

$$g_{1i}(\sigma^2) = (1 - \gamma_i)\sigma_v^2, \quad (5.1)$$

$$g_{2i}(\sigma^2) = (\bar{\mathbf{X}}_i - \gamma_i\bar{\mathbf{x}}_i)'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\bar{\mathbf{X}}_i - \gamma_i\bar{\mathbf{x}}_i), \quad (5.2)$$

and

$$g_{3i}(\sigma^2) = n_i^{-2}(\sigma_v^2 + \sigma_e^2 n_i^{-1})^{-3}[\sigma_e^4\,\text{var}(\tilde{\sigma}_v^2) + \sigma_v^4\,\text{var}(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_v^2\,\text{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2)] \quad (5.3)$$

are all of order $O(t^{-1})$. Furthermore, under normality of $\{v_i\}$ and $\{e_{ij}\}$

$$\text{var}(\tilde{\sigma}_e^2) = 2(n - t - k + \lambda)^{-1}\sigma_e^4, \quad (5.4)$$

$$\text{var}(\tilde{\sigma}_v^2) = 2n_*^{-2}[(n - t - k + \lambda)^{-1} \times (t - \lambda)(n - k)\sigma_e^4 + 2n_*\sigma_e^2\sigma_v^2 + n_{**}\sigma_v^4], \quad (5.5)$$

and

$$\text{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2) = -(t - \lambda)n_*^{-1}\text{var}(\tilde{\sigma}_e^2), \quad (5.6)$$

where $n_{**} = \text{tr}(\mathbf{MZZ}')^2$ with $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (see Battese and Fuller 1981).

Estimators of $g_{2i}(\sigma^2)$ and $g_{3i}(\sigma^2)$ are simply given by $g_{2i}(\hat{\sigma}^2)$ and $g_{3i}(\hat{\sigma}^2)$ correct to $O_p(t^{-1})$, since $\hat{\sigma}^2$ is a consistent estimator of σ^2. Nevertheless, $g_{1i}(\hat{\sigma}^2)$ is not the correct estimator of $g_{1i}(\sigma^2)$ to the desired order of approximation, because its bias is of order $O(t^{-1})$. A correct estimator of $g_{1i}(\sigma^2)$ is obtained by adjusting $g_{1i}(\hat{\sigma}^2)$ for its bias to $O(t^{-1})$, which is obtained by making a Taylor expansion of $g_{1i}(\hat{\sigma}^2)$ around σ^2 and then taking its expectation. After considerable algebraic simplification, we obtain $Eg_{1i}(\hat{\sigma}^2) - g_{1i}(\sigma^2) = -g_{3i}(\sigma^2) + o(t^{-1})$. Therefore, $g_{1i}(\hat{\sigma}^2) + g_{3i}(\hat{\sigma}^2)$ is correct to $O_p(t^{-1})$ in estimating $g_{1i}(\sigma^2)$. It now follows that an estimator of the MSE approximation with expectation correct to $O(t^{-1})$ is given by

$$mse[t_i(\hat{\sigma}^2, \mathbf{y})] = g_{1i}(\hat{\sigma}^2) + g_{2i}(\hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}^2). \quad (5.7)$$

Similarly, for the random regression coefficient model, $mse[t_i(\hat{\sigma}^2, \mathbf{y})]$ is given by (5.7), with

$$g_{1i}(\sigma^2) = \bar{X}_i^2(1 - \bar{\gamma}_i)\sigma_v^2, \quad (5.8)$$

$$g_{2i}(\sigma^2) = \bar{X}_i^2(1 - \bar{\gamma}_i)^2(\Sigma\bar{\gamma}_i)^{-1}\sigma_v^2, \quad (5.9)$$

and

$$g_{3i}(\sigma^2) = \bar{X}_i^2(\Sigma_j x_{ij}^2)^{-2}[\sigma_v^2 + \sigma_e^2(\Sigma_j x_{ij}^2)^{-1}]^{-3} \times [\sigma_e^4\text{var}(\tilde{\sigma}_v^2) + \sigma_v^4\text{var}(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_v^2\text{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2)]. \quad (5.10)$$

Furthermore, under normality of $\{v_i\}$ and $\{e_{ij}\}$

$$\text{var}(\tilde{\sigma}_e^2) = 2(n - t)^{-1}\sigma_e^4, \quad (5.11)$$

$$\text{var}(\tilde{\sigma}_v^2) = 2\tilde{n}_*^{-2}[(n - 1)(t - 1)(n - t)^{-1}\sigma_e^4 + 2\tilde{n}_*\sigma_e^2\sigma_v^2 + \tilde{n}_{**}\sigma_v^4], \quad (5.12)$$

and

$$\text{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2) = -(t - 1)\tilde{n}_*^{-1}\text{var}(\tilde{\sigma}_e^2), \quad (5.13)$$

where $\tilde{n}_{**} = \text{tr}(\mathbf{MZZ}')^2$.

It follows from (4.12) that an estimator of MSE approximation for finite populations is given by

$$mse[t_i^F(\hat{\sigma}^2, \mathbf{y})] = (1 - f_i)^2[mse(t_i^*(\hat{\sigma}^2, \mathbf{y})) + N_i^{-1}(1 - f_i)^{-1}\hat{\sigma}_e^2], \quad (5.14)$$

for the nested-error regression and random regression coefficient models. Here, $mse(t_i^*(\hat{\sigma}^2, \mathbf{y}))$ is given by (5.7), with $\bar{X}_i$ replaced by $\bar{x}_i^*$.

Turning to the Fay–Herriot model, we get

$$mse[t_i(\hat{A}, \bar{y})] = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}), \quad (5.15)$$

where

$$g_{1i}(A) = AD_i(A + D_i)^{-1}, \quad (5.16)$$

$$g_{2i}(A) = D_i^2(A + D_i)^{-2}\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i, \quad (5.17)$$

and

$$g_{3i}(A) = D_i^2(A + D_i)^{-3}\text{var}(\bar{A}). \quad (5.18)$$

Furthermore, under normality of $\{v_i\}$ and $\{e_i\}$

$$\text{var}(\bar{A}) \doteq 2t^{-1}[A^2 + 2A\Sigma D_i/t + \Sigma D_i^2/t], \quad (5.19)$$

where the neglected terms in the approximation are of lower order than $O(t^{-1})$. Morris (1983) proposed empirical Bayesian confidence intervals for the mean μ_i in the equal-variance situation with $D_i = D$ through a slight generalization of his previous result (Morris 1981) for the special case of $\mu_i = \mu + v_i$. His intervals are given by $\hat{\mu}_i \pm z_{\alpha/2}s_i$, where $z_{\alpha/2}$ is the upper-$\alpha/2$ point of the $N(0, 1)$ distribution, $\hat{\mu}_i$ is the empirical Bayes estimator essentially equivalent to $t_i(\hat{A}, \bar{y})$, and $s_i^2 \doteq g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2D^2(\hat{A} + D)^{-2}t^{-1}(y_i - \mathbf{x}_i\hat{\beta})^2$, neglecting terms of order $o(t^{-1})$ in his formula for s_i^2. It is interesting to note that s_i^2 reduces to $g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + g_{3i}(\hat{A})$ to the order of approximation if we replace $(y_i - \mathbf{x}_i'\hat{\beta})^2$ with its average value, $t^{-1}\Sigma(y_i - \mathbf{x}_i'\hat{\beta})^2$. This estimator is equivalent to the estimator of MSE obtained through our approach if $g_{1i}(\hat{A})$ is not adjusted for its bias to $O(t^{-1})$. Morris (1983) gave a heuristic extension of s_i^2 to the case of unequal D_i, but the comparison of his formula with (5.15) is not clear (although the first two terms agree).

Following Morris (1983), we propose $t_i(\hat{\sigma}^2, \mathbf{y}) \pm z_{\alpha/2}[mse(t_i(\hat{\sigma}^2, \mathbf{y}))]^{1/2}$ as confidence intervals for the nested-error regression and random regression coefficient models, and $t_i(\hat{A}, \bar{y}) \pm z_{\alpha/2}[mse(t_i(\hat{A}, \bar{y}))]^{1/2}$ for the Fay–Herriot model, for nominal level $1 - \alpha$. As $t \to \infty$, $\hat{\sigma}^2$ and $\hat{A}$ converge in probability to σ^2 and A, respectively, and hence it follows from Slutsky's theorem (e.g., see Bickel and Doksum 1977, p. 461) that the coverage probability of these intervals tends to the desired level, $1 - \alpha$, provided that the random errors $\{v_i\}$ and $\{e_{ij}\}$ (or $\{e_i\}$ in the Fay–Herriot model) are both normally distributed. Morris noted this asymptotic property for his intervals as well.

6. RESULTS OF A MONTE CARLO STUDY

A Monte Carlo study under the nested-error regression model with one auxiliary variable, $y_{ij} = \alpha + \beta x_{ij} + v_i + e_{ij}$, was conducted to study the efficiency of two-stage estimators, the accuracy of the second-order approximation to MSE, and the relative bias of estimators of MSE. We used the values $\alpha = 5.5$, $\beta = .388$, $\sigma_e^2 = 292$, and $\sigma_v^2 = 64$, which were obtained by Battese et al. (1988) as estimates from some data (y_{ij}, x_{ij}), with y_{ij} equal to corn acres and x_{ij} equal to the number of pixels of corn for the jth sample segment of county i (in Iowa) ($j = 1, \ldots, n_i$; $i = 1, \ldots, 12$). In their data set, $n_i = 1$ for three of the counties. We pooled these three counties, and we increased the number of small areas, t, to 20 from 10 by duplicating $(x_{ij}, n_i, \bar{X}_i)$ (reported by Battese and Fuller). We then generated 10,000 independent sets of normal variates e_{ij} ($j = 1, \ldots, n_i$; $i = 1, \ldots, 20$) and v_i ($i = 1, \ldots, 20$) from $N(0, \sigma_e^2 = 292)$ and $N(0, \sigma_v^2 = 64)$. Using the given x_{ij} values, we then obtained 10,000 sets of $\{y_{ij}: j = 1, \ldots, n_i; i = 1, \ldots, 20\}$ from the model $y_{ij} = 5.5 + .388x_{ij} + v_i + e_{ij}$. Monte Carlo values of MSE$[t_i(\hat{\sigma}^2, \mathbf{y})]$, $E\ mse[t_i(\hat{\sigma}^2, \mathbf{y})]$, and so forth were computed from the 10,000 data sets.

Similarly, independent data sets were generated from the following nonnormal distributions: double-exponential (symmetric, long-tailed), uniform (short-tailed), and exponential (positively skewed), such that v_i and e_{ij} have means 0 and variances 64 and 292, respectively.

A brief summary of the Monte Carlo results is given in the following, but the details can be found in Prasad and Rao (1986).

6.1 Efficiency of Two-Stage Estimators

The relative efficiency of the two-stage estimator $t_i(\hat{\sigma}^2, \mathbf{y})$ under normal errors $\{v_i\}$ and $\{e_{ij}\}$ ranged from 123% to 184% with respect to the regression synthetic estimator $\bar{y}_i(\text{syn}) = \hat{\alpha} + \hat{\beta}\bar{x}_i$, and from 142% to 174% with respect to the approximately unbiased regression estimator $\bar{y}_i(\text{reg}) = \bar{y}_i + \hat{\beta}(\bar{X}_i - \bar{x}_i)$, where $\hat{\alpha}$ and $\hat{\beta}$ are the ordinary least squares estimators of α and β, respectively. The relative efficiency with respect to $\bar{y}_i(\text{syn})$ increases as n_i increases from 2 to 6, whereas the relative efficiency with respect to $\bar{y}_i(\text{reg})$ exhibits an opposite trend; that is, it decreases as n_i increases.

6.2 Accuracy of the Second-Order Approximation to MSE

The relative error (RE) of the second-order approximation to MSE, averaged over small areas having the same n_i value, is small ($\leq 2\%$) under normal errors $\{v_i\}$ and $\{e_{ij}\}$, and it is not large under deviations from normality for $\{e_{ij}\}$ only ($< 7\%$). Nevertheless, it leads to considerable overstatement of MSE; RE ranges from 11% to 19% when both errors are generated from the exponential distribution. In addition, it is not quite satisfactory when both errors are double-exponential, with RE ranging from 6% to 14%. Under uniform distribution for both errors, the approximation leads to a slight understatement.

The accuracy of the approximation depends on the negligibility of the cross-product term, $2E[t_i(\hat{\sigma}^2, \mathbf{y}) - t_i(\sigma^2, \mathbf{y})][t_i(\sigma^2, \mathbf{y}) - \mu_i]$/MSE, which is exactly 0 under normality. The value of the cross-product term ranged, as n_i increases from 2 to 6, from -5% to -15% under exponential distributions and -4% to -8% under double-exponential distributions, compared with -1% to -4% under exponential distributions for $\{e_{ij}\}$ only and -1% to -2.5% under double-exponential distributions for $\{e_{ij}\}$ only; that is, the approximation is satisfactory when the random effects $\{v_i\}$ are approximately normal.

6.3 Relative Bias of Estimators of MSE

The relative bias of the normality-based estimator of MSE, $mse[t_i(\hat{\sigma}^2, \mathbf{y})]$, averaged over small areas having the same n_i value, is small (< 7%) when both $\{u_i\}$ and $\{e_{ij}\}$ are normal or uniform, or when $\{v_i\}$ are normal and $\{e_{ij}\}$ are uniform. Nevertheless, as n_i increases from 2 to 6 it ranges from 3% to 16% under double-exponential distributions and from 2% to 19% under exponential distributions, compared with 2% to 11% under double-exponential distributions for $\{e_{ij}\}$ only, and 1% to 14% under exponential distributions for $\{e_{ij}\}$ only. The customary estimator of MSE, $g_1(\hat{\sigma}^2) + g_2(\hat{\sigma}^2)$, which ignores the uncertainty in the estimator $\hat{\sigma}^2$, leads to severe underestimation of true MSE.

Overall, the Monte Carlo study indicated that the two-stage estimators lead to considerable gain in efficiency over the customary regression synthetic estimator or the approximately unbiased regression estimator. In addition, (5.7) is a reliable estimator of MSE for approximately normal or short-tailed distributions of $\{v_i\}$ and $\{e_{ij}\}$, and to a lesser extent it is reliable for long-tailed or positively skewed distributions of $\{e_{ij}\}$ and approximately normal $\{v_i\}$.

APPENDIX: PROOFS

A.1 Order of Approximation to the MSE of the Two-Stage Estimator

We consider the general mixed model (3.1) with $\mathbf{y} = \text{col}_{1\le i\le I}(\mathbf{y}_i)$, $\mathbf{X} = \text{col}_{1\le i\le I}(\mathbf{X}_i)$, $\mathbf{Z} = \text{diag}_{1\le i\le I}(\mathbf{Z}_i)$, $\mathbf{v} = \text{col}_{1\le i\le I}(\mathbf{v}_i)$, and $\mathbf{e} = \text{col}_{1\le i\le I}(\mathbf{e}_i)$, where $\mathbf{y}_i$ and $\mathbf{e}_i$ are $n_i \times 1$ random vectors, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are (respectively) $n_i \times k$ and $n_i \times b_i$ matrices of known constants, $\mathbf{v}_i$ is a $b_i \times 1$ random vector, and $n = \Sigma n_i$ is the total sample size. Furthermore, $\mathbf{e}_i$ and $\mathbf{v}_i$ are such that $\mathbf{R} = \text{diag}_{1\le i\le I}(\mathbf{R}_i)$, $\mathbf{G} = \text{diag}_{1\le i\le I}(\mathbf{G}_i)$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, where $\mathbf{R}_i$ and $\mathbf{G}_i$ are (respectively) $n_i \times n_i$ and $b_i \times b_i$ matrices depending on $\boldsymbol{\theta}$. In addition, $\mu_i = \mathbf{k}_i'\boldsymbol{\beta} + \mathbf{m}_i'\mathbf{v}$ with $\mathbf{m}_i = \text{col}_{1\le i\le I}(\delta_{il}\mathbf{m}_{il})$, where $\delta_{il} = 1$ if $i = l$ and $\delta_{il} = 0$ if $i \ne l$, $\mathbf{k}_i$ and $\mathbf{m}_{il}$ are (respectively) $p \times 1$ and $b_i \times 1$ vectors of known constants, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters.

We assume the following regularity conditions:

1. The elements of $\mathbf{X}$ and $\mathbf{Z}$ are uniformly bounded such that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = O(t^{-1})$.
2. $\sup_{i\ge 1} n_i = \Delta_1 < \infty$ and $\sup_{i\ge 1} b_i = \Delta_2 < \infty$.
3. The elements of R_i and G_i are uniformly bounded and differentiable with respect to $\boldsymbol{\theta}$.
4. $\tilde{\theta}_j = \mathbf{y}'\mathbf{C}_j\mathbf{y}$ is a translation-invariant unbiased estimator of θ_j, where $\mathbf{C}_j$ is of the form $\mathbf{C}_j = \text{diag}_{1\le i\le I}[O(t^{-1})]_{n_i\times n_i} + [O(t^{-2})]_{n\times n}$. Here $[O(t^{-s})]_{m\times m}$ denotes an $m \times m$ matrix with elements uniformly of order $O(t^{-s})$.

We now state and prove our main theorem, which shows that the order of neglected terms in Approximations (4.3) and (4.4) to $E[\mathbf{d}(\boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2$ is $o(t^{-1})$.

Theorem A.1. Suppose that the regularity conditions 1–4 hold. Then, under normality for the random errors in the model (3.1),

$$E[\mathbf{d}(\boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 = \text{tr}[(\nabla\mathbf{b}_i')\mathbf{V}(\nabla\mathbf{b}_i')'E(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'] + o(t^{-1}), \quad \text{(A.1)}$$

where $\nabla\mathbf{b}_i' = \text{col}_{1\le j\le p}(\partial\mathbf{b}_i'/\partial\theta_j)$ and $\mathbf{b}_i' = \mathbf{m}_i'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}$.

The proof of Theorem A.1 involves several lemmas.

Lemma A.1. Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be nonstochastic matrices of order n, and $\mathbf{u} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then,

$$E[\mathbf{u}(\mathbf{u}'\mathbf{A}_s\mathbf{u})\mathbf{u}'] = (\text{tr } \mathbf{A}_s\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}\mathbf{A}_s\boldsymbol{\Sigma}, \qquad s = 1, 2,$$

$$E\left[\prod_{s=1}^{2}(\mathbf{u}'\mathbf{A}_s\mathbf{u})\right] = 2 \text{ tr } \mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2\boldsymbol{\Sigma} + (\text{tr } \mathbf{A}_1\boldsymbol{\Sigma})(\text{tr } \mathbf{A}_2\boldsymbol{\Sigma}),$$

and

$$E\left[\mathbf{u}\left(\prod_{s=1}^{2}\mathbf{u}'\mathbf{A}_s\mathbf{u}\right)\mathbf{u}'\right] = (\text{tr } \mathbf{A}_1\boldsymbol{\Sigma})(\text{tr } \mathbf{A}_2\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2(\text{tr } A_1\boldsymbol{\Sigma})\boldsymbol{\Sigma}\mathbf{A}_2\boldsymbol{\Sigma} + 2(\text{tr } \mathbf{A}_2\boldsymbol{\Sigma})\boldsymbol{\Sigma}\mathbf{A}_1\boldsymbol{\Sigma} + 2(\text{tr } \mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 4\boldsymbol{\Sigma}\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2\boldsymbol{\Sigma} + 4\boldsymbol{\Sigma}\mathbf{A}_2\boldsymbol{\Sigma}\mathbf{A}_1\boldsymbol{\Sigma}.$$

Proof. Proof is obtained through direct extension of the results of Srivastava and Tiwari (1976) for $N_n(\mathbf{0}, \mathbf{I})$.

Lemma A.2. Let $\mathbf{u} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$, $z_j = \boldsymbol{\lambda}_j'\mathbf{u}$, and $q_j = \mathbf{u}'\mathbf{A}_j\mathbf{u}$ $(j = 1, \ldots, p)$, where $\boldsymbol{\lambda}_j$ and $\mathbf{A}_j$ are nonstochastic of order $n \times 1$ and $n \times n$, respectively. Then,

$$E[\mathbf{z}'(\mathbf{q} - E\mathbf{q})]^2 = \text{tr } \boldsymbol{\Sigma}_z\boldsymbol{\Sigma}_q + 4\sum_{j=1}^{p}\sum_{l=1}^{p}\{\boldsymbol{\lambda}_j'\boldsymbol{\Sigma}\mathbf{A}_j\boldsymbol{\Sigma}\mathbf{A}_l\boldsymbol{\Sigma}\boldsymbol{\lambda}_l + \boldsymbol{\lambda}_j'\boldsymbol{\Sigma}\mathbf{A}_l\boldsymbol{\Sigma}\mathbf{A}_j\boldsymbol{\Sigma}\boldsymbol{\lambda}_l\}, \quad \text{(A.2)}$$

where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)'$, $\mathbf{q} = (q_1, \ldots, q_p)'$, and $\boldsymbol{\Sigma}_z$ and $\boldsymbol{\Sigma}_q$ are the covariance matrices of $\mathbf{z}$ and $\mathbf{q}$, respectively.

Proof. The result of (A.2) follows from Lemma A.1, after noting the following: $Eq_i = \text{tr } \mathbf{A}_i\boldsymbol{\Sigma}$, $\text{cov}(q_i, q_l) = 2 \text{ tr } \mathbf{A}_i\boldsymbol{\Sigma}\mathbf{A}_l\boldsymbol{\Sigma}$, and

$$E[z_iz_l(q_i - Eq_i)(q_l - Eq_l)] = \boldsymbol{\lambda}_i'E[\mathbf{u}(\mathbf{u}'\mathbf{A}_i\mathbf{u}\mathbf{u}'\mathbf{A}_l\mathbf{u})\mathbf{u}']\boldsymbol{\lambda}_l - (Eq_i)\boldsymbol{\lambda}_i'E[\mathbf{u}(\mathbf{u}'\mathbf{A}_l\mathbf{u})\mathbf{u}']\boldsymbol{\lambda}_l - (Eq_l)\boldsymbol{\lambda}_i'E[\mathbf{u}(\mathbf{u}'\mathbf{A}_i\mathbf{u})\mathbf{u}']\boldsymbol{\lambda}_l + (Eq_i)(Eq_l)\boldsymbol{\lambda}_i'\boldsymbol{\Sigma}\boldsymbol{\lambda}_l.$$

Lemma A.3. Let (a) $\boldsymbol{\Sigma}_{n\times n} = \text{diag}_{1\le i\le I}(\boldsymbol{\Sigma}_i)$, (b) $\mathbf{C}_{n\times n} = \text{diag}_{1\le i\le I}[O(t^{-1})]_{n_i\times n_i} + [O(t^{-2})]_{n\times n}$, (c) $\mathbf{r} = \text{col}_{1\le k\le I}\text{col}_{1\le j\le n_k}(O(t^{-1}))$, and (d) $\mathbf{s}_i = \text{col}_{1\le k\le I}\text{col}_{1\le j\le n_k}(\delta_{ik}O(1))$, where $\boldsymbol{\Sigma}_i$ is an $n_i \times n_i$ matrix with bounded elements. Then, the following results hold: (e) $\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma} = [O(t^{-2})]_{n\times n}$; (f) $\mathbf{s}_i'\boldsymbol{\Sigma}\mathbf{s}_i = O(1)$; (g) $(\mathbf{r} + \mathbf{s}_i)'\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma}(\mathbf{r} + \mathbf{s}_i) = O(t^{-2})$.

Proof. By computing the indicated products using (a)–(d), results (e)–(g) are obtained.

Lemma A.4. Under the regularity conditions 1–3, letting $d_i^{(j)}(\boldsymbol{\theta})$ be the jth element of $\mathbf{d}_i(\boldsymbol{\theta}) = \partial t_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, we obtain

$$\text{cov}[d_i^{(j)}(\boldsymbol{\theta}), d_i^{(l)}(\boldsymbol{\theta})] = [\partial\mathbf{b}_i'/\partial\theta_j]\mathbf{V}[\partial\mathbf{b}_i'/\partial\theta_l]' + O(t^{-1}). \quad \text{(A.3)}$$

Proof. Using $\partial(\mathbf{AB})/\partial\boldsymbol{\theta} = (\partial\mathbf{A}/\partial\boldsymbol{\theta})\mathbf{B} + A(\partial\mathbf{B}/\partial\boldsymbol{\theta})$ and $\partial\mathbf{B}^{-1}/\partial\boldsymbol{\theta} = -B^{-1}(\partial\mathbf{B}/\partial\boldsymbol{\theta})\mathbf{B}^{-1}$, we can write $d_i^{(j)}(\boldsymbol{\theta}) = (\mathbf{f}_i(j)' + \partial\mathbf{b}_i'/\partial\theta_j)\mathbf{u}$, where $\mathbf{u} = \mathbf{Z}\mathbf{v} + \mathbf{e}$, and

$$\mathbf{f}_i(j)' = (\mathbf{1}' - \mathbf{m}_i'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \times \mathbf{X}'(\partial\mathbf{V}^{-1}/\partial\theta_j)\mathbf{A} - (\partial\mathbf{b}_i'/\partial\theta_j)\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1},$$

with $\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. Therefore,

$$\text{cov}[d_i^{(j)}(\boldsymbol{\theta}, \mathbf{y}), d_i^{(l)}(\boldsymbol{\theta}, \mathbf{y})] = (\mathbf{f}_i'(j) + \partial\mathbf{b}_i'/\partial\theta_j)\mathbf{V}(\mathbf{f}_i'(l) + \partial\mathbf{b}_i'/\partial\theta_l)'. \quad \text{(A.4)}$$

Now, from the regularity conditions 1 and 2 it follows that the elements of $\mathbf{f}_i(j)$ are $O(t^{-1})$ and $\partial\mathbf{b}_i'/\partial\theta_j = [\text{col}_{1\le i\le I}(\delta_{il}\mathbf{O}(1))]'$, by exploiting the block-diagonal structure of $\mathbf{G}$ and $\mathbf{Z}$ and $\mathbf{V}^{-1}$, and then by verifying that $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' = [O(t^{-1})]_{n\times n}$, $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\partial\mathbf{V}^{-1}/\partial\theta_j\mathbf{A} = [O(t^{-1})]_{n\times n}$, and $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} = [O(t^{-1})]_{n\times n}$. Hence by Lemma A.3 (f) the

desired result (A.3) follows from (A.4), noting that $\mathbf{V}$ is the form $\boldsymbol{\Sigma}$ of Lemma A.3.

Proof of Theorem A.1. The desired result (A.1) is obtained from (A.4) by first showing that

$$E[\mathbf{d}_i(\theta, \mathbf{y})'(\tilde{\theta} - \theta)]^2 = \operatorname{tr}[\mathbf{A}_i(\theta)E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)'] + o(t^{-1}) \tag{A.5}$$

and then using Lemma A.4 to replace $\mathbf{A}_i(\theta)$ with $(\nabla \mathbf{b}_i')\mathbf{V}(\nabla \mathbf{b}_i')'$.

Using Lemma A.2 with $\boldsymbol{\lambda}_i' = \mathbf{f}_i(j)' + \partial \mathbf{b}_i'/\partial\theta_j = \boldsymbol{\lambda}_j(i)'$ (say), $\mathbf{z} = \mathbf{d}_i(\theta)$, $\boldsymbol{\Sigma} = \mathbf{V}$, $\mathbf{A}_j = \mathbf{C}_j$, and $\mathbf{q} = \tilde{\theta}$, we get

$$E[\mathbf{d}_i(\theta, \mathbf{y})'(\tilde{\theta} - \theta)]^2 = \operatorname{tr}[A_i(\theta)E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)']$$

$$+ 4\sum_{j=1}^{p}\sum_{l=1}^{p}[\boldsymbol{\lambda}_j(i)'\mathbf{V}\mathbf{C}_l\mathbf{V}\mathbf{C}_l\mathbf{V}\boldsymbol{\lambda}_l(i) + \boldsymbol{\lambda}_j(i)'\mathbf{V}\mathbf{C}_l\mathbf{V}\mathbf{C}_j\mathbf{V}\boldsymbol{\lambda}_l(i)]. \tag{A.6}$$

Now, noting that $\boldsymbol{\lambda}_j(i)$ and $\mathbf{C}_j$ are of the form $\mathbf{r} + \mathbf{s}_i$ and $\mathbf{C}$ of Lemma A.3, it follows that the last two terms of (A.6) are of order $o(t^{-1})$. Hence (A.5) is true.

A.2 Order of Taylor Series Approximation (4.2) for the Fay–Herriot Model

We now show that the order of neglected terms in the Taylor series approximation (4.2), under the Fay–Herriot model, is $o(t^{-1})$.

Theorem A.2. Let the variances D_i satisfy $0 < D_L \le D_i \le D_U < \infty$ for all i, and let $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ be such that $\max_{i\le t} h_i = O(t^{-1})$. Then,

$$E[t_i(\hat{A}, \mathbf{y}) - t_i(A, \bar{\mathbf{y}})]^2 = E[(\hat{A} - A)\partial t_i(A, \bar{\mathbf{y}})/\partial A]^2 + o(t^{-1}). \tag{A.7}$$

The proof of Theorem A.2 requires the following two lemmas.

Lemma A.5. $E(\hat{A} - A)^{2s} = O(t^{-s})$ $(s \ge 1)$.

Proof. We write $\hat{A} - A$ as

$$\hat{A} - A = (t - k)^{-1}[\Sigma(\bar{y}_i - \mathbf{x}_i'\beta)^2 - (\bar{\beta} - \beta)'\mathbf{X}'\mathbf{X}(\bar{\beta} - \beta) - \Sigma(A + D_i)(1 - h_i)]$$

$$= (t - k)^{-1}[\Sigma U_i(A + D_i) - T\Sigma(A + D_i)h_i], \tag{A.8}$$

where $\bar{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{y}}$. Furthermore, $U_i = (\bar{y}_i - \mathbf{x}_i'\beta)^2/(A + D_i) - 1 = \tilde{U}_i - 1$, with $E(U_i) = 0$ and where the $\tilde{U}_i$'s are independent χ_1^2 variables, and $T = (\bar{\beta} - \beta')(\mathbf{X}'\mathbf{X})^{-1}(\bar{\beta} - \beta)/\Sigma(A + D_i)h_i$ with $E(T) = 0$ and $\tilde{T} = \sum_1^k \lambda_i V_i/\sum_i^k \lambda_i$, where the V_i's are independent χ_1^2 variables, the λ_i's are the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})$, and $\Sigma\lambda_i = \Sigma(A + D_i)h_i$. The eigenvalues λ_i are bounded, since $\lambda_{\max} \le A + D_U$.

It follows from (A.8) that

$$E(\hat{A} - A)^2 \le 2(t - k)^{-2}[E(\Sigma U_i(A + D_i))^2 + ET^2(\Sigma(A + D_i)h_i)^2] = O(t^{-1}),$$

since $ET^2 = O(1)$, $\Sigma(A + D_i)h_i \asymp O(1)$, and $E(\Sigma U_i(A + D_i))^2 = \Sigma EU_i^2(A + D_i)^2 \le (A + D_U)^2 tE(U_1^2) = O(t)$. Similarly,

$$E(\hat{A} - A)^{2s} \le 2^s(t - k)^{-2s}[E(\Sigma U_i(A + D_i))^{2s} + ET^{2s}(\Sigma(A + D_i)h_i)^{2s}]$$

$$= O(t^{-s}), \qquad s \ge 2,$$

noting that $E(\Sigma U_i(A + D_i))^{2s} = O(t^s)$, since $E(U_i) = 0$.

Lemma A.6. $E(\hat{A} - A)^{2s} = O(t^{-s})$ $(s \ge 1)$ and $E[\hat{A} - \tilde{A}]^4 = O(t^{-4})$.

Proof. We have

$$\Pr(\hat{A} \le 0) = \Pr(\hat{A} - A \le -A) \le \Pr(|\hat{A} - A| \ge A) \le E(\hat{A} - A)^{2l}/A^{2l}$$

by Chebyshev's inequality. Hence $\Pr(\hat{A} \le 0) = O(t^{-l})$ for any desired l, by Lemma A.5.

It now follows that

$$E(\hat{A} - A)^{2s} = E[(\hat{A} - A)^{2s}|\hat{A} \le 0]\Pr(\hat{A} \le 0) + E[(\hat{A} - A)^{2s}|\hat{A} > 0]\Pr(\hat{A} > 0)$$

$$\le A^{2s}\Pr(\hat{A} \le 0) + E(\hat{A} - A)^{2s} = O(t^{-s}),$$

choosing $l = s$ and using Lemma A.5. Furthermore, writing $(\hat{A} - \tilde{A})^4 = \tilde{A}^4 W$, where $W = 1$ if $\tilde{A} \le 0$ and $W = 0$ if $\tilde{A} > 0$, from the Cauchy–Schwarz inequality we get $E[\hat{A} - \tilde{A}]^4 \le [E\tilde{A}^8\Pr(\tilde{A} < 0)]^{1/2} = O(t^{-4})$, by choosing $\Pr(\tilde{A} \le 0) = O(t^{-8})$ and noting that $E\tilde{A}^8 \le 8[E(\hat{A} - A)^8 + A^8] = O(1)$.

Proof of Theorem A.2. By Taylor series expansion of $t_i(\hat{A}, \bar{\mathbf{y}})$ about the point A, we have

$$t_i(\hat{A}, \bar{\mathbf{y}}) - t_i(A, \bar{\mathbf{y}}) = (\hat{A} - A)\partial t_i(A, \bar{\mathbf{y}})/\partial A + \tfrac{1}{2}(\hat{A} - A)^2\partial^2 t_i(A^*, \bar{\mathbf{y}})/\partial A^{*2},$$

where $|A^* - A| < |\hat{A} - A|$. Hence

$$E[t_i(\hat{A}, \bar{\mathbf{y}}) - t_i(A, \bar{\mathbf{y}})]^2 = E[(\hat{A} - A)\partial t_i(A, \bar{\mathbf{y}})/\partial A]^2 + R_1 + R_2,$$

where

$$R_1 = E[(\hat{A} - A)\partial t_i(A, \bar{\mathbf{y}})/\partial A][(\hat{A} - A)^2\partial^2 t_i(A^*, \bar{\mathbf{y}})/\partial A^{*2}]$$

and $R_2 = E[(\hat{A} - A)^4\{\partial^2 t_i(A^*, \bar{\mathbf{y}})/\partial A^{*2}\}^2]$.

We first show that R_2 is $o(t^{-1})$. We have

$$\partial t_i(A, \bar{\mathbf{y}})/\partial A = D_i(A + D_i)^{-1}\partial(\mathbf{x}_i'\tilde{\beta})/\partial A + D_i(A + D_i)^{-2}(\bar{y}_i - \mathbf{x}_i'\tilde{\beta}),$$

where $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\bar{\mathbf{y}})$ and $\partial(\mathbf{x}_i'\tilde{\beta})/\partial A = -\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-2}\tilde{\mathbf{u}})$, with $\tilde{\mathbf{u}} = \bar{\mathbf{y}} - \mathbf{X}\tilde{\beta}$. The following matrix result is used repeatedly in the subsequent steps of the proof: If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular such that $\mathbf{z}'\mathbf{A}\mathbf{z} > \mathbf{z}'\mathbf{B}\mathbf{z}$, then $\mathbf{z}'\mathbf{A}^{-1}\mathbf{z} < \mathbf{z}'\mathbf{B}^{-1}\mathbf{z}$ for every $\mathbf{z} \ne 0$ [e.g., see Graybill 1969, theorem 12.2.14, result (5)]. Using this result and the Cauchy–Schwarz inequality, and noting that $\tilde{\mathbf{u}}'\mathbf{V}^{-1}\tilde{\mathbf{u}} \le \mathbf{u}'\mathbf{V}^{-1}\mathbf{u}$, where $\mathbf{u} = \bar{\mathbf{y}} - \mathbf{X}\beta$, we get

$$|\partial(\mathbf{x}_i'\tilde{\beta})/\partial A| \le (\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i)^{1/2}(\tilde{\mathbf{u}}'\mathbf{V}^{-2}\tilde{\mathbf{u}})^{1/2}$$

$$\le (\max_i h_i)^{1/2}(A + D_U)^{1/2}D_L^{-3/2}(\mathbf{u}'\mathbf{u})^{1/2}. \tag{A.9}$$

Hence

$$|\partial t_i(A, \bar{\mathbf{y}})/\partial A| \le |\partial(\mathbf{x}_i'\tilde{\beta})/\partial A| + D_L^{-1}|\bar{y}_i - \mathbf{x}_i'\tilde{\beta}|, \tag{A.10}$$

where $|\partial(\mathbf{x}_i'\tilde{\beta})/\partial A|$ is given by (A.9).

Turning to the second derivative of $t_i(A, \bar{\mathbf{y}})$, we get

$$\partial^2 t_i(A, \bar{\mathbf{y}})/\partial A^2 = D_i(A + D_i)^{-1}\partial^2(\mathbf{x}_i'\tilde{\beta})/\partial A^2 - 2D_i(A + D_i)^{-2}\partial(\mathbf{x}_i'\tilde{\beta})/\partial A - 2D_i(A + D_i)^{-3}(\bar{y}_i - \mathbf{x}_i'\tilde{\beta}),$$

where

$$\partial^2(\mathbf{x}_i'\tilde{\beta})/\partial A^2 = 2\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-3}\tilde{\mathbf{u}} - 2\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\tilde{\mathbf{u}}.$$

Hence, applying the Cauchy–Schwarz inequality and the matrix result to the previous two terms, after some simplification, we get

$$|\partial^2(\mathbf{x}_i'\tilde{\beta})/\partial A^2| \le 4(\max_i h_i)^{1/2}(A + D_U)^{1/2}D_L^{-5/2}(\mathbf{u}'\mathbf{u})^{1/2}, \tag{A.11}$$

noting that $\tilde{\mathbf{u}}'\mathbf{V}^{-2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\tilde{\mathbf{u}} \le \tilde{\mathbf{u}}'\mathbf{V}^{-3}\tilde{\mathbf{u}}$, since $\mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1/2}$ is a symmetric, idempotent matrix with eigenvalues 1 or 0. Hence using (A.9) we obtain

$$|\partial^2 t_i(A, \bar{\mathbf{y}})/\partial A^2| \le 6(\max_i h_i)^{1/2}(A + D_U)^{1/2}D_L^{-5/2}(\mathbf{u}'\mathbf{u})^{1/2} + 2D_L^{-2}|\bar{y}_i - \mathbf{x}_i'\tilde{\beta}|.$$

Therefore,

$$[\partial^2 l_i(A, \bar{y})/\partial A^2]^2 \le c_1(A + D_U)((1/t)\Sigma u_i^2) + c_2(\bar{y}_i - \mathbf{x}_i'\tilde{\beta})^2, \quad \text{(A.12)}$$

where c_1 and c_2 are $O(1)$, noting that $t \max_i h_i = O(1)$. Furthermore,

$$(\bar{y}_i - \mathbf{x}_i'\tilde{\beta})^2 \le 2[u_i^2 + (\mathbf{x}_i'(\tilde{\beta} - \beta))^2] \le 2[u_i^2 + (A + D_U)(t \max_i h_i) D_L^{-1}((1/t)\Sigma u_i^2)],$$

by using the Cauchy–Schwarz inequality on $(\mathbf{x}_i'(\tilde{\beta} - \beta))^2$, where $u_i = y_i - \mathbf{x}_i'\beta$. Therefore, from (A.12) we get $R_2 \le E[\tilde{c}_1(\hat{A} - A)^4(A^* + D_U)((1/t)\Sigma u_i^2) + 2c_2(\hat{A} - A)^4 u_i^2]$, where $\tilde{c}_1$ is $O(1)$.

Now,

$$|E(\hat{A} - A)^4(A^* + D_U)((1/t)\Sigma u_i^2)| < E|\hat{A} - A|^5((1/t)\Sigma u_i^2) + (A + D_U)E(\hat{A} - A)^4((1/t)\Sigma u_i^2), \quad \text{(A.13)}$$

since $|A^* - A| < |\hat{A} - A|$. Therefore, by Lemma A.6 and the Cauchy–Schwarz inequality the first term on the right side of (A.13) is $o(t^{-1})$, noting that $E((1/t)\Sigma u_i^2)^2$ is $O(1)$ because the u_i's are independent $N(0, A + D_i)$. Similarly, $E(\hat{A} - A)^4 u_i^2 = o(t^{-1})$ because $E(u_i^4)$ is $O(1)$. Hence R_2 is $o(t^{-1})$.

By again appealing to the Cauchy–Schwarz inequality, it follows that R_1 is $o(t^{-1})$ as well, noting that $E[(\hat{A} - A)\partial l_i(A, \bar{y})/\partial A]^2$ is $O(t^{-1})$ using (A.10).

Finally, the result $E[(\hat{A} - A)\partial l_i(A, \bar{y})/\partial A]^2 = E[(\tilde{A} - A)\partial l_i(A, \bar{y})/\partial A]^2 + o(t^{-1})$ follows by using the result $E(\hat{A} - \tilde{A})^4 = O(t^{-4})$ and the Cauchy–Schwarz inequality, and writing $\hat{A} - A = A - \tilde{A} + \hat{A} - A$.

A.3 Order of Approximation to the Estimator of MSE for the Fay–Herriot Model

We now show that the approximation to the estimator of MSE, (5.15), has expectation correct to $O(t^{-1})$ under the Fay–Herriot model.

Theorem A.3. Let the conditions of Theorem A.2 hold. Then,

$$E[g_{1i}(\hat{A})] = g_{1i}(A) - g_{3i}(A) + o(t^{-1}), \quad \text{(A.14)}$$

$$E[g_{2i}(\hat{A})] = g_{2i}(A) + o(t^{-1}), \quad \text{(A.15)}$$

and

$$E[g_{3i}(\hat{A})] = g_{3i}(A) + o(t^{-1}). \quad \text{(A.16)}$$

Proof. We first consider $g_{1i}(\hat{A}) = \hat{A}D_i(\hat{A} + D_i)^{-1}$. By Taylor series expansion of $g_{1i}(\hat{A})$ around A, we get

$$g_{1i}(\hat{A}) = g_{1i}(A) + (\hat{A} - A)g_{1i}'(A) + \tfrac{1}{2}(\hat{A} - A)^2 g_{1i}''(A) + \tfrac{1}{2}(\hat{A} - A)^2[g_{1i}''(A^*) - g_{1i}''(A)], \quad \text{(A.17)}$$

where $|A^* - A| < |\hat{A} - A|$. Now, noting that $E(\hat{A} - A) = E(\hat{A} - \tilde{A})$ and $E|\hat{A} - \tilde{A}| = o(t^{-l})$, by using $\Pr(\tilde{A} \le 0) = O(t^{-l})$ for any desired l (Lemma A.6) we get $|E(\hat{A} - A)g_{1i}'(A)| \le g_{1i}'(A)E|\hat{A} - A| = o(t^{-1})$, since $g_{1i}'(A) = D_i^2(A + D_i)^{-2} < 1$.

In addition, $E(\hat{A} - A)^2 = E(\hat{A} - A)^2 + o(t^{-1}) = \text{var}(\hat{A}) + o(t^{-1})$, from Lemma A.6. Hence

$$E[\tfrac{1}{2}(\hat{A} - A)^2 g_{1i}''(A)] = \tfrac{1}{2}\,\text{var}(\hat{A})g_{1i}''(A) + o(t^{-1}) = -g_{3i}(A) + o(t^{-1}),$$

noting that $g_{1i}''(A) = -2D_i^2(A + D_i)^{-3}$. For the last term of (A.17), we have

$$|g_{1i}''(A^*) - g_{1i}''(A)| = 2D_i^2(A + D_i)^{-3}(A^* + D_i)^{-3}|(A^* + D_i)^3 - (A + D_i)^3| \le 2D_L^{-4}[|A^* - A|^3 + 3(A + D_U)(A^* - A)^2 + 3(A + D_U)^2|A^* - A|] < 2D_L^{-4}[|\hat{A} - A|^3 + 3(A + D_U)(\hat{A} - A)^2 + 3(A + D_U)^2|\hat{A} - A|].$$

Hence

$$E(\hat{A} - A)^2|g_{1i}''(A^*) - g_{1i}''(A)| \le 2D_L^{-4}[E|\hat{A} - A|^5 + 3(A + D_U)E(\hat{A} - A)^4 + 3(A + D_U)^2 E|\hat{A} - A|^3] = o(t^{-1}),$$

noting that $E(\hat{A} - A)^{2a} = O(t^{-a})$ so that $E|\hat{A} - A|^3 \le [E(\hat{A} - A)^2 E(\hat{A} - A)^4]^{1/2} = o(t^{-1})$ and $E|\hat{A} - A|^5 \le [E(\hat{A} - A)^4 E(\hat{A} - A)^6]^{1/2} = o(t^{-1})$. Hence (A.14) is established.

Turning to $g_{2i}(\hat{A})$, we have

$$g_{2i}(\hat{A}) - g_{2i}(A) = D_i^2(\hat{A} + D_i)^{-2}[\mathbf{x}_i'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{x}_i - \mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i] + D_i^2\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i[(\hat{A} + D_i)^{-2} - (A + D_i)^{-2}]. \quad \text{(A.18)}$$

By Taylor series expansion around A, we get

$$I_1 = \mathbf{x}_i'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{x}_i - \mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i = (\hat{A} - A)[\partial\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i/\partial A]_{A=A^*} = -(\hat{A} - A)\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{*-2}\mathbf{X})(\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\mathbf{x}_i,$$

where $\mathbf{V}^*$ is the value of $\mathbf{V}$ when $A = A^*$. Using the previous matrix result, we get $|I_1| < D_L^{-1}|\hat{A} - A|\,|\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\mathbf{x}_i|$ or

$$E|I_1| < D_L^{-1}(\max_i h_i)E[|\hat{A} - A|(A^* + D_U)] \le \frac{1}{t} D_L^{-1}(t \max_i h_i)[E(\hat{A} - A)^2 E(A^* + D_U)^2]^{1/2} = o(t^{-1}),$$

noting that $t \max_i h_i = O(1)$, $E(\hat{A} - A)^2 = O(t^{-1})$, and $E(A^* + D_U)^2 \le 2[E(\hat{A} - A)^2 + (A + D_U)^2] = O(1)$. Similarly,

$$E|(\hat{A} + D_i)^{-2} - (A + D_i)^{-2}| < D_L^{-4}[E(\hat{A} - A)^2 + 2(A + D_U)E|\hat{A} - A|] = O(t^{-1/2}),$$

noting that $[E|\hat{A} - A|]^2 \le E(\hat{A} - A)^2$. In addition, $\mathbf{x}_i'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_i \le (1/t)(A + D_U)(t \max_i h_i) = O(t^{-1})$. It now follows from (A.18) that $E|g_{2i}(\hat{A}) - g_{2i}(A)| = o(t^{-1})$, so (A.15) is established.

Finally, turning to $g_{3i}(\hat{A})$ we have

$$g_{3i}(\hat{A}) - g_{3i}(A) = (2/t)D_i^2(A + D_i)^{-3}[(\hat{A} - A)^2 + 2(A + \bar{D})(\hat{A} - A)] + (2/t)D_i^2(A^2 + 2A\bar{D} + \Sigma D_i^2/t) \times [(A + D_i)^{-3} - (\hat{A} + D_i)^{-3}] = I_3 + I_4, \quad \text{(A.19)}$$

say, where $\bar{D} = \Sigma D_i/t$. Hence $E|I_2| < (2/t)D_L^{-1}[E(\hat{A} - A)^2 + 2(A + D_U)E|\hat{A} - A|] = o(t^{-1})$ and

$$E|I_3| < (2/t)D_L^{-4}(A^2 + 2AD_U + D_U^2)[E|\hat{A} - A|^3 + 3(A + D_U)E(\hat{A} - A)^2 + 3(A + D_U)E|\hat{A} - A|] = o(t^{-1}),$$

noting that $E|\hat{A} - A| = O(t^{-1/2})$. It now follows from (A.19) that $E|g_{3i}(\hat{A}) - g_{3i}(A)| = o(t^{-1})$, so (A.16) is established.

[*Received October 1986. Revised January 1989.*]

REFERENCES

Battese, G. E., and Fuller, W. A. (1981), "Prediction of County Crop Areas Using Survey and Satellite Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 500–505.

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.

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981), "Estimation in Covariance Component Models," *Journal of the American Statistical Association*, 76, 341–353.

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

Ghosh, M., and Lahiri, P. (1987), "Robust Empirical Bayes Estimation of Means From Stratified Samples," *Journal of the American Statistical Association*, 82, 1153–1162.

Ghosh, M., and Meeden, G. (1986), "Empirical Bayes Estimation in Finite Population Sampling," *Journal of the American Statistical Association*, 81, 1058–1062.

Graybill, F. A. (1969), *Introduction to Matrices With Applications in Statistics*, Belmont, CA: Wadsworth.

Harville, D. A. (1977), "Maximum Likelihood Approach to Variance Component Estimation and Related Problems," *Journal of the American Statistical Association*, 72, 320–340.

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

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–862.

Morris, C. E. (1981), "Parametric Empirical Bayes Confidence Intervals," in *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and C. F. J. Wu, New York: Academic Press, pp. 25–50.

——— (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–59.

Peixoto, J. L., and Harville, D. A. (1986), "Comparisons of Alternative Predictors Under the Balanced One-Way Random Model," *Journal of the American Statistical Association*, 81, 431–436.

Prasad, N. G. N., and Rao, J. N. K. (1986), "On the Estimation of Mean Square Error of Small Area Predictors," Technical Report 97, Carleton University, Laboratory for Research in Statistics and Probability, Ottawa.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Srivastava, V. K., and Tiwari, R. (1976), "Evaluation of Expectation of Products of Stochastic Matrices," *Scandinavian Journal of Statistics*, 3, 135–138.

The Annals of Statistics
1991, Vol. 19, No. 4, 1748–1770

BAYESIAN PREDICTION IN LINEAR MODELS: APPLICATIONS TO SMALL AREA ESTIMATION[1]

BY GAURI SANKAR DATTA AND MALAY GHOSH

University of Georgia and University of Florida

This paper introduces a hierarchical Bayes (HB) approach for prediction in general mixed linear models. The results find application in small area estimation. Our model unifies and extends a number of models previously considered in this area. Computational formulas for obtaining the Bayes predictors and their standard errors are given in the general case. The methods are applied to two actual data sets. Also, in a special case, the HB predictors are shown to possess some interesting frequentist properties.

1. Introduction. It has been some time now that the government agencies in the United States, Canada and elsewhere have recognized the importance of small area estimation. Estimation of this type is particularly well suited in a setting that involves several areas (or strata) with a small number of samples available from each individual stratum. The estimates of the parameters of interest (like the mean, variance, etc.) for these areas can profitably "borrow strength" from other neighboring areas.

The appropriateness of model-based inference for small area estimation is widely recognized. We may refer to Fay and Herriot (1979), Ghosh and Meeden (1986), Ghosh and Lahiri (1987), Battese, Harter and Fuller (1988), Prasad and Rao (1990), Choudhry and Rao (1988), Royall (1978) and Lui and Cumberland (1989), among others. The methods that have usually been proposed use either a variance components approach or an empirical Bayes (EB) approach, although the distinction between the two is often superfluous [Harville (1988, 1990)]. Both these procedures use certain mixed linear models for prediction purposes. First, assuming the variance components are known, certain best linear unbiased predictors (BLUPs) or EB predictors are obtained for the unknown parameters of interest. Then the unknown variance components are estimated typically by Henderson's method of fitting of constants or the restricted maximum likelihood (REML) method, and the resulting estimated BLUPs (also referred to as empirical BLUPs) are used for final prediction.

Received March, 1989; revised November 1990.

[1]The second author's research was partially supported by NSF Grants DMS-87-01814 and DMS-89-01334. Part of this work was completed when the second author was an ASA Senior Research Fellow at the Bureau of the Census and the Bureau of Labor Statistics, while the first author was an ASA/NSF/Census Research Associate.

AMS 1980 *subject classifications*. 62D05, 62F11, 62F15, 62J99.

Key words and phrases. Hierarchical Bayes, empirical Bayes, mixed linear models, best linear unbiased prediction, best unbiased prediction, small area estimation, nested error regression model, random regression coefficients model, two-stage sampling, elliptically symmetric distributions.

Although the above approach is usually quite satisfactory for point prediction, it is very difficult to estimate the standard errors associated with these predictors. This is primarily due to the lack of closed-form expressions for the mean squared errors (MSEs) of the estimated BLUPs. Kackar and Harville (1984) suggested an approximation to the MSEs [see also Harville (1985, 1988, 1990) and Harville and Jeske (1989)]. Prasad and Rao (1990) proposed estimates of these approximate MSEs in three specific mixed linear models. The work of Prasad and Rao (1990) suggests that their approximations work well when the number of small areas is sufficiently large. It is not clear though how these approximations fare for a small or even a moderately large number of strata.

Ghosh and Lahiri (1989) proposed a hierarchical Bayes (HB) procedure as an alternative to the estimated BLUP or the EB procedure. In a HB procedure, if one uses the posterior mean for estimating the parameter of interest, then a natural estimate of the standard error associated with this estimator is the posterior s.d. The estimate, though often complicated, can be found exactly via numerical integration without any approximation.

The model considered by Ghosh and Lahiri (1989) was, however, only a special case of the so-called "nested error regression model." A similar model was considered by Stroud (1987), but his general analysis was performed only for the balanced case, that is, when the number of samples was the same for each stratum. Battese, Harter and Fuller (1988) first considered the nested error regression model in the context of small area estimation and performed a variance components analysis.

The objective of this article is to present a unified Bayesian prediction theory for mixed linear models with particular emphasis on small area estimation. A general Bayesian normal theory model is presented in Section 2 which can be regarded as an extension of the HB ideas of Lindley and Smith (1972) to prediction. Most of the models considered by earlier authors can be regarded as special cases of our model, and certain specific illustrations are provided. Also, in this section, we have provided in a very general framework the posterior distribution as well as the resulting posterior means and variances of the unobserved population units given the sampled units. The proof of the main result of this section is given in the Appendix. For nonnormal HB analysis, one may refer to Albert (1988) or Morris (1988).

In Section 3, we discuss the computational issues related to the estimation of parameters of interest with particular emphasis on the estimation of population means simultaneously for several small areas. Closed-form expressions cannot usually be obtained for the posterior means and s.d.'s of such parameters, and numerical integration becomes a necessity. For very high dimensional integrals, direct numerical integration is often unreliable, and sometimes even impossible to execute, and some of the recently advocated Monte Carlo integration techniques may be of help. We shall indicate in Section 3 how the Gibbs sampling technique introduced by Geman and Geman (1984), and more recently popularized by Gelfand and Smith (1990), works in some important special cases of our general framework. The related substitu-

tion sampling algorithm of Tanner and Wong (1987) and the traditional importance sampling technique will also be discussed very briefly.

However, in small dimensions, it is often easier to perform direct numerical integration than to use any Monte Carlo numerical integration method. For instance, if the integrand is a very complicated function and cannot be approximated very accurately by a simple smooth function, the importance sampling technique can at best result in a slow convergence of the desired integral. The Gibbs sampling is usually very slow, and for evaluation of small dimensional integrals, any simplicity of this approach cannot adequately compensate for the enormous computing time needed for the method's successful execution.

For the sake of illustration of our methods, we have thus used in Section 4 direct numerical integration methods for data analysis. Two examples are considered in this paper. The first example given in Section 4.1 requires numerical evaluation of two-dimensional integrals, while the second given in Section 4.2 requires evaluation of one-dimensional integrals. The data set considered in the first example pertains to the Patterns of Care Studies, a study involving the quality of treatment received by cancer patients having radiation therapy as the primary mode of treatment. The present data form a subset of a much larger data set analyzed in Calvin and Sedransk (1991). We have considered a stratified finite population from which samples are drawn in two stages using simple random sampling at each stage. The HB estimator of the population mean is compared with an EB estimator proposed in Ghosh and Lahiri (1988), a design unbiased estimator given in Cochran (1977), page 303, an expansion estimator, a ratio type estimator and another estimator proposed in Royall (1976). The HB estimator has the smallest average mean squared error among these six and the improvement over all but the EB estimator is quite substantial.

The second example is related to the prediction of areas under corn and soybeans for 12 counties in North Central Iowa. The problem was originally considered by Battese, Harter and Fuller (1988) using a variance components method. We have used this example to illustrate how a naive EB approach can sometimes grossly underestimate the associated standard error of an EB estimator. In this particular example, the posterior s.d.'s as obtained by us are slightly smaller than the ones of Battese, Harter and Fuller.

In Section 5, we have considered a special case of the general HB model and have provided the posterior distribution of the unobserved population units given the sampled units. In this special case, the HB predictors of the linear parameters of interest are shown to be the best within the class of all linear unbiased predictors under the assumption of finiteness of second moments. For a class of spherically symmetric distributions including but not limited to the normal, the HB predictors are shown to be optimal within the class of all unbiased predictors. Optimality properties of this type extend the earlier work of Henderson (1963) and others on the prediction of real-valued parameters to the prediction of vector-valued parameters. The proof of the main result of Section 2 is deferred to the Appendix.

2. The description and analysis of the HB model. Consider the following Bayesian model:

(A) Conditional on $\mathbf{b} = (b_1, \ldots, b_p)^T$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t)^T$ and r, let

$$\mathbf{Y} \sim N\big(\mathbf{Xb}, r^{-1}(\boldsymbol{\Psi} + \mathbf{ZD}(\boldsymbol{\lambda})\mathbf{Z}^T)\big),$$

where $\mathbf{Y}$ is $N \times 1$.

(B) $\mathbf{B}$, R and $\boldsymbol{\Lambda}$ have a certain joint prior distribution proper or improper.

Stage (A) of the model can be identified as a general mixed linear model. To see this, write

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zv} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{e}$ and $\mathbf{v}$ are mutually independent, with $\mathbf{e} \sim N(\mathbf{0}, r^{-1}\boldsymbol{\Psi})$ and $\mathbf{v} \sim N(\mathbf{0}, r^{-1}\mathbf{D}(\boldsymbol{\lambda}))$, where $\mathbf{e}$ is $N \times 1$ and $\mathbf{v}$ is $q \times 1$; in the above $\mathbf{X}$ $(N \times p)$ and $\mathbf{Z}$ $(N \times q)$ are known design matrices, $\boldsymbol{\Psi}$ is a known positive definite (p.d.) matrix, while $\mathbf{D}(\boldsymbol{\lambda})$ $(q \times q)$ is a p.d. matrix which is structurally known except possibly for some unknown $\boldsymbol{\lambda}$. In the examples to follow, $\boldsymbol{\lambda}$ involves the ratios of the variance components. Sometimes we will denote $\mathbf{D}(\boldsymbol{\lambda})$ by $\mathbf{D}$ when $\boldsymbol{\lambda}$ is known.

In the context of small area estimation, partition $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{e}$, and rewrite the model given in (2.1) as

$$\begin{pmatrix}\mathbf{Y}^{(1)}\\ \mathbf{Y}^{(2)}\end{pmatrix} = \begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}\mathbf{b} + \begin{pmatrix}\mathbf{Z}^{(1)}\\ \mathbf{Z}^{(2)}\end{pmatrix}\mathbf{v} + \begin{pmatrix}\mathbf{e}^{(1)}\\ \mathbf{e}^{(2)}\end{pmatrix}, \tag{2.2}$$

where $\mathbf{Y}^{(1)}$ and $\mathbf{e}^{(1)}$ are $n \times 1$, $\mathbf{X}^{(1)}$ is $n \times p$ and $\mathbf{Z}^{(1)}$ is $n \times q$. Also, $\mathbf{Y}^{(2)}$ and $\mathbf{e}^{(2)}$ are $(N - n) \times 1$, $\mathbf{X}^{(2)}$ is $(N - n) \times p$ and $\mathbf{Z}^{(2)}$ is $(N - n) \times q$. We assume for simplicity that $\operatorname{rank}(\mathbf{X}^{(1)}) = p$.

In the above $\mathbf{Y}^{(1)}$ is the vector of sampled units from m small areas, while $\mathbf{Y}^{(2)}$ is the vector of unsampled units. It is possible to partition $\mathbf{Y}^{(1)T}$ into $\mathbf{Y}^{(1)T} = (\mathbf{Y}_1^{(1)T}, \ldots, \mathbf{Y}_m^{(1)T})$, where $\mathbf{Y}_i^{(1)}(n_i \times 1)$ is the vector of sampled units for the ith small area. Similarly, $\mathbf{Y}^{(2)T}$ can be partitioned as $\mathbf{Y}^{(2)T} = (\mathbf{Y}_1^{(2)T}, \ldots, \mathbf{Y}_m^{(2)T})$, where $\mathbf{Y}_i^{(2)}((N_i - n_i) \times 1)$ is the vector of unsampled units for the ith small area.

Following the model-based approach in survey sampling, one of the primary objectives of this paper is to find the conditional (predictive) distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. The analysis will be done in two stages. In the latter part of this section, we derive the predictive distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$ putting independent uniform prior distributions on $\mathbf{B}$ and gamma distributions on $R, \Lambda_1 R, \ldots, \Lambda_t R$.

Before finding the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$, we identify some of the existing models introduced for small area estimation by several authors as special cases of (2.2). In what follows, we shall use the notation $\mathbf{I}_u$ for an identity matrix of order u, $\mathbf{1}_u$ for a u-component column vector with each element equal to 1 and $\mathbf{J}_u = \mathbf{1}_u\mathbf{1}_u^T$. Also, let $\operatorname{col}_{1 \le i \le p}(\mathbf{B}_i)$ denote the matrix $(\mathbf{B}_1^T, \ldots, \mathbf{B}_p^T)^T$ and let $\oplus_{i=1}^p \mathbf{A}_i$ denote the matrix $\begin{bmatrix}\mathbf{A}_1 & \cdots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{A}_p\end{bmatrix}$.

First, consider the nested error regression model

$$Y_{ij} = \mathbf{x}_{ij}^T\mathbf{b} + v_i + e_{ij}, \qquad j = 1, \ldots, N_i, i = 1, \ldots, m. \tag{2.3}$$

The model was considered by Battese, Harter and Fuller (1988). They assumed the v_i's and e_{ij}'s to be mutually independent with v_i's iid $\mathrm{N}(0, (\lambda r)^{-1})$, and e_{ij}'s iid $N(0, r^{-1})$. In this case, $\mathbf{X}^{(1)} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{1 \le j \le n_i}(\mathbf{x}_{ij}^T))$, $\mathbf{X}^{(2)} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{n_i+1 \le j \le N_i}(\mathbf{x}_{ij}^T))$, $\mathbf{Z}^{(1)} = \oplus_{i=1}^m \mathbf{1}_{n_i}$ and $\mathbf{Z}^{(2)} = \oplus_{i=1}^m \mathbf{1}_{N_i - n_i}$, $\boldsymbol{\Psi} = \mathbf{I}_N$, $t = 1$, $\boldsymbol{\lambda} = \lambda$ and $\mathbf{D}(\boldsymbol{\lambda}) = \lambda^{-1}\mathbf{I}_m$. In the further special case of Ghosh and Lahiri (1989), $\mathbf{x}_{ij} = \mathbf{x}_i$ for every $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Note that $\lambda = V(e_{ij})/V(v_i)$, a ratio of the variance components.

The random regression coefficients model of Dempster, Rubin and Tsutakawa (1981) [see also Prasad and Rao (1990)] is also a special case of ours. In this setup, $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\boldsymbol{\Psi}$ and $\mathbf{D}(\boldsymbol{\lambda})$ are the same as in the nested error regression model, but

$$\mathbf{Z}^{(1)} = \bigoplus_{i=1}^m \left[\mathrm{col}_{1 \le j \le n_i}\mathbf{x}_{ij}^T\right], \qquad \mathbf{Z}^{(2)} = \bigotimes_{i=1}^m \left[\mathrm{col}_{n_i+1 \le j \le N_i}\mathbf{x}_{ij}^T\right].$$

The models given in Choudhry and Rao (1988) are special cases of our general model as well.

It is possible also to include certain cross-classification models as special cases of our general linear model. For example, suppose there are m small areas labeled $1, \ldots, m$. Within each small area, units are further classified into c subgroups (socioeconomic class, age, etc.) labeled $1, \ldots, c$. The cell sizes N_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, c$, are assumed to be known. Let Y_{ijk}, $k = 1, \ldots, N_{ij}$, denote the measurement on the kth individual in the (i, j)th cell. Conditional on $\mathbf{b}$, r and $\boldsymbol{\lambda}$, suppose

$$\begin{aligned} Y_{ijk} &= \mathbf{x}_{ij}^T\mathbf{b} + \tau_i + \eta_j + \gamma_{ij} + e_{ijk}, \\ &\qquad k = 1, \ldots, N_{ij}, i = 1, \ldots, m, j = 1, \ldots, c, \end{aligned} \tag{2.4}$$

with τ_i's, η_j's, γ_{ij}'s and e_{ijk}'s mutually independent with e_{ijk}'s iid $N(0, r^{-1})$, γ_{ij}'s iid $N(0, (\lambda_3 r)^{-1})$, η_j's iid $N(0, (\lambda_2 r)^{-1})$ and τ_i's iid $N(0, (\lambda_1 r)^{-1})$. Special cases of this model have been considered by several authors. Lui and Cumberland (1989) [also Royall (1978)] considered a model where τ_i's and γ_{ij}'s are degenerate at zeros. Also, they assumed the variance ratio λ_2 to be known in deriving their estimators and did not address the issue of unknown λ_2 appropriately.

Next we show that the two-stage sampling model with covariates and m strata is a special case of our general linear model. Suppose that the ith stratum contains L_i primary units. Suppose also that the jth primary unit within the ith stratum contains N_{ij} subunits. Let Y_{ijk} denote the value of the characteristic of interest for the kth subunit within the jth primary unit from the ith stratum ($k = 1, \ldots, N_{ij}$, $j = 1, \ldots, L_i$, $i = 1, \ldots, m$). From the ith stratum, a sample of l_i primary units is taken. For the jth selected primary unit within the ith stratum, a sample of n_{ij} subunits are selected. Without

loss of generality, the sample values are denoted by Y_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, l_i$, $i = 1, \ldots, m$.

Assume conditional on $\mathbf{b}$, r and $\boldsymbol{\lambda}$:

$$Y_{ijk} = \mathbf{x}_{ij}^T \mathbf{b} + \xi_i + \eta_{ij} + e_{ijk}, \qquad k = 1, \ldots, N_{ij},\ j = 1, \ldots, L_i,\ i = 1, \ldots, m, \tag{2.5}$$

where ξ_i's, η_{ij}'s and e_{ijk}'s are mutually independent with ξ_i's iid $N(0, (\lambda_1 r)^{-1})$, η_{ij}'s iid $N(0, (\lambda_2 r)^{-1})$, e_{ijk}'s iid $N(0, r^{-1})$. Let

$$\mathbf{Y}^{(1)} = \underset{1 \le i \le m}{\operatorname{col}} \left[\underset{1 \le j \le l_i}{\operatorname{col}} \left\{ \underset{1 \le k \le n_{ij}}{\operatorname{col}} (Y_{ijk}) \right\} \right],$$

$$\mathbf{Y}^{(2)} = \underset{1 \le i \le m}{\operatorname{col}} \left[\underset{1 \le j \le L_i}{\operatorname{col}} \left\{ \underset{u_{ij} \le k \le N_{ij}}{\operatorname{col}} (Y_{ijk}) \right\} \right],$$

$$u_{ij} = 1 + n_{ij} I_{[j \le l_i]},\ i = 1, \ldots, m.$$

$$\mathbf{v} = \left(\mathbf{s}^T \mathbf{w}_1^T \mathbf{w}_2^T\right)^T, \qquad \mathbf{s} = \underset{1 \le i \le m}{\operatorname{col}} (\xi_i), \qquad \mathbf{w}_1 = \underset{1 \le i \le m}{\operatorname{col}} \left(\underset{1 \le j \le l_i}{\operatorname{col}} (\eta_{ij}) \right)$$

and

$$\mathbf{w}_2 = \underset{1 \le i \le m}{\operatorname{col}} \left(\underset{l_i + 1 \le j \le L_i}{\operatorname{col}} (\eta_{ij}) \right).$$

Also, let $\mathbf{e}^{(i)}$ be defined similarly as $\mathbf{Y}^{(i)}$, $i = 1, 2$. Then (2.5) can be written as (2.2) with appropriately defined $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Note that here $t = 2$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, $\boldsymbol{\Psi} = \mathbf{I}_N$, $\mathbf{D}(\boldsymbol{\lambda}) = \operatorname{Diag}(\lambda_1^{-1} \mathbf{I}_m, \lambda_2^{-1} \mathbf{I}_L.)$ with $N = \sum_{i=1}^m \sum_{j=1}^{L_i} N_{ij}$. The ideas can be extended directly to multistage sampling. We may mention here that Bayesian analysis for two-stage sampling was introduced first by Scott and Smith (1969) in a much simpler framework. A multistage analog of their work was provided by Malec and Sedransk (1985).

Next, in this section, we provide the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. The following nomenclature will be used to label certain known distributions. A random variable Z is said to have a Gamma(α, β) distribution if it has pdf $f(z) = [\exp(-\alpha z) \alpha^\beta z^{\beta - 1} / \Gamma(\beta)] I_{[z > 0]}$. A random vector $\mathbf{T} = (T_1, \ldots, T_p)^T$ is said to have a multivariate t-distribution with location parameter $\boldsymbol{\mu}$, scale parameter $\boldsymbol{\Phi}$ and degrees of freedom ν if it has pdf

$$g(\mathbf{t}) \propto |\boldsymbol{\Phi}|^{-1/2} \left[\nu + (\mathbf{t} - \boldsymbol{\mu})^T \boldsymbol{\Phi}^{-1} (\mathbf{t} - \boldsymbol{\mu}) \right]^{-(\nu + p)/2} \tag{2.6}$$

[see Zellner (1971) page 383, or Press (1972) page 136]. Assume $\nu > 2$. Then $E(\mathbf{T}) = \boldsymbol{\mu}$, $V(\mathbf{T}) = (\nu/(\nu - 2))\boldsymbol{\Phi}$.

We assume condition (A) given at the beginning of this section. In stage (B) of the model, it is assumed that

$$\mathbf{B}, R, \Lambda_1 R, \ldots, \Lambda_t R \text{ are independently distributed} \tag{2.7}$$

with $\mathbf{B} \sim$ uniform(R^p), $R \sim$ Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$, $a_0 \ge 0$, $g_0 \ge 0$, $\Lambda_i R \sim$ Gamma$(\frac{1}{2}a_i, \frac{1}{2}g_i)$, $i = 1, \ldots, t$, with $a_i > 0$, $g_i \ge 0$, $i = 1, \ldots, t$. In this way, some improper gamma distributions are included as a possibility in our prior.

Before stating the main result of this section we need to introduce additional notation. We write $\Sigma \equiv \Sigma(\lambda) = \Psi + \mathbf{Z}\mathbf{D}(\lambda)\mathbf{Z}^T$, partition Σ into $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and define $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

Also, let

$$\mathbf{K} = \Sigma_{11}^{-1} - \Sigma_{11}^{-1}\mathbf{X}^{(1)}\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}, \tag{2.8}$$

$$\mathbf{M} = \Sigma_{21}\mathbf{K} + \mathbf{X}^{(2)}\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}, \tag{2.9}$$

$$\begin{aligned}\mathbf{G} = \Sigma_{22.1} &+ \left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1} \\ &\times\left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{T}.\end{aligned} \tag{2.10}$$

The posterior distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ is given in the following theorem in two steps.

THEOREM 1. *Consider the model given in* (2.1) [*or* (2.2)] *and* (2.7). *Assume that* $n + \Sigma_{i=0}^{t} g_i - p > 2$. *Then, conditional on* $\Lambda = \lambda$ *and* $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$, $\mathbf{Y}^{(2)}$ *is distributed as multivariate-t with degrees of freedom* $n + \Sigma_{i=0}^{t} g_i - p$, *location parameter* $\mathbf{M}\mathbf{y}^{(1)}$ *and scale parameter*

$$\left(n + \sum_{i=0}^{t} g_i - p\right)^{-1}\left[a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right]\mathbf{G}.$$

Also, the conditional distribution of Λ *given* $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ *has pdf*

$$\begin{aligned}f(\lambda|\mathbf{y}^{(1)}) \propto |\Sigma_{11}|^{-1/2}&\left|\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right|^{-1/2}\left[\prod_{i=1}^{t}\lambda_i^{gi/2-1}\right] \\ &\times\left[a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right]^{-[n+\Sigma_{i=0}^{t} g_i - p]/2}.\end{aligned} \tag{2.11}$$

The proof of Theorem 1 is deferred to the Appendix. Using the moments of a multivariate-t distribution, it follows now that if $n + \Sigma_{i=0}^{t} g_i > p + 2$, then

$$E\left[\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}\right] = E\left(\mathbf{M}|\mathbf{y}^{(1)}\right)\mathbf{y}^{(1)}, \tag{2.12}$$

$$\begin{aligned}V\left[\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}\right] = V\left(\mathbf{M}\mathbf{y}^{(1)}|\mathbf{y}^{(1)}\right) &+ \left(n + \sum_{i=0}^{t} g_i - p - 2\right)^{-1} \\ &\times E\left[\left\{a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right\}\mathbf{G}\middle|\mathbf{y}^{(1)}\right].\end{aligned} \tag{2.13}$$

Using (2.12) and (2.13), it is possible to find the posterior means and variances of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \mathbf{A}\mathbf{Y}^{(1)} + \mathbf{C}\mathbf{Y}^{(2)}$, where $\mathbf{A}$ and $\mathbf{C}$ are known matrices. The Bayes estimate of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ under any quadratic loss is its posterior

mean, and is given by

$$\mathbf{e}_B(\mathbf{y}^{(1)}) = \left[\mathbf{A} + \mathbf{C}E(\mathbf{M}|\mathbf{y}^{(1)})\right]\mathbf{y}^{(1)}, \tag{2.14}$$

using (2.12). Similarly, using (2.13), one may obtain

$$V\left[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{y}^{(1)}\right] = \mathbf{C}V(\mathbf{Y}^{(2)}|\mathbf{y}^{(1)})\mathbf{C}^T. \tag{2.15}$$

Note that when $\mathbf{A} = \oplus_{i=1}^m \mathbf{1}_{n_i}^T$ and $\mathbf{C} = \oplus_{i=1}^m \mathbf{1}_{N_i - n_i}^T$, $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ reduces to the vector of population totals for the m small areas. Computational issues related to the simultaneous estimation of several small area totals will be addressed in Section 3.

3. Numerical computations. It is evident from Theorem 1 that the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$ cannot usually be obtained analytically because of the complicated posterior pdf of Λ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ [see (2.11)]. As mentioned in the Introduction, Monte Carlo numerical integration is a distinct possibility, particularly when the dimension of λ is large. One may think of the importance sampling method as a natural candidate for such purposes. To implement such a procedure, we write $f(\lambda|\mathbf{y}^{(1)})$ given in (2.11) as $f(\lambda|\mathbf{y}^{(1)}) = ck(\lambda, \mathbf{y}^{(1)})$, where the norming constant c has to be numerically evaluated. Now, for any real-valued function $h(\lambda)$,

$$\int_0^\infty \cdots \int_0^\infty h(\lambda) f(\lambda|\mathbf{y}^{(1)})\, d\lambda = \frac{\int_0^\infty \cdots \int_0^\infty h(\lambda)\{k(\lambda, \mathbf{y}^{(1)})/g(\lambda|\mathbf{y}^{(1)})\}g(\lambda|\mathbf{y}^{(1)})\, d\lambda}{\int_0^\infty \cdots \int_0^\infty \{k(\lambda, \mathbf{y}^{(1)})/g(\lambda|\mathbf{y}^{(1)})\}g(\lambda|\mathbf{y}^{(1)})\, d\lambda},$$

where $g(\lambda|\mathbf{y}^{(1)})$ is some "standard" pdf from which a random sample can easily be generated. Hence $\int_0^\infty \cdots \int_0^\infty h(\lambda) f(\lambda|\mathbf{y}^{(1)})\, d\lambda$ can be approximated by

$$\frac{\sum_{i=1}^s h(\lambda^{(i)})\{k(\lambda^{(i)}, \mathbf{y}^{(1)})/g(\lambda^{(i)}|\mathbf{y}^{(1)})\}}{\sum_{i=1}^s k(\lambda^{(i)}, \mathbf{y}^{(1)})/g(\lambda^{(i)}|\mathbf{y}^{(1)})},$$

where the number of replicates is very large, and $\lambda^{(i)}$'s are generated from $g(\lambda|\mathbf{y}^{(1)})$.

Unfortunately, finding $g(\lambda|\mathbf{y}^{(1)})$ in the present context can be quite formidable. Even when λ is one-dimensional, $f(\lambda|\mathbf{y}^{(1)})$ may turn out to be multimodal, and thus defy any simple approximation. One such example appears in Ghosh and Rao (1991). In such circumstances, it is natural to seek other Monte Carlo integration methods.

The recently advertised Gibbs sampler bears some interesting promise, at least in the special case when $\Psi = \mathbf{I}_N$ and $\mathbf{D}(\lambda) = \text{Diag}(\lambda_1^{-1}\mathbf{I}_{q_1}, \ldots, \lambda_t^{-1}\mathbf{I}_{q_t})$, where $\sum_{i=1}^t q_i = q$. We shall write $W_i = R\Lambda_i$, and correspondingly $w_i = r\lambda_i$, $i = 1, \ldots, t$. We assign a uniform (R^p) prior for $\mathbf{B}$, a Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$ prior for R and Gamma$(\frac{1}{2}a_i, \frac{1}{2}g_i)$ priors for the W_i's, where $\mathbf{B}, R, W_1, \ldots, W_t$ are all independently distributed.

We shall write $\mathbf{v}^T = (\mathbf{v}_1^T, \ldots, \mathbf{v}_t^T)$, where $\mathbf{v}_i$ has dimension q_i. Based on the model introduced at the beginning of Section 2, the joint pdf of $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{B}, \mathbf{v}, R, W_1, \ldots, W_t$ is

$$
\begin{aligned}
&f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, r, w_1, \ldots, w_t) \\
&\quad \propto r^{n/2} \exp\Big[-\tfrac{1}{2} r \big\| \mathbf{y}^{(1)} - \mathbf{X}^{(1)}\mathbf{b} - \mathbf{Z}^{(1)}\mathbf{v} \big\|^2\Big] r^{(N-n)/2} \\
&\qquad \times \exp\Big[-\tfrac{1}{2} r \big\| \mathbf{y}^{(2)} - \mathbf{X}^{(2)}\mathbf{b} - \mathbf{Z}^{(2)}\mathbf{v} \big\|^2\Big] \\
&\qquad \times \prod_{i=1}^{t} \Big\{ w_i^{q_i/2} \exp\big(-\tfrac{1}{2} w_i \|\mathbf{v}_i\|^2\big)\Big\} \exp\big(-\tfrac{1}{2} a_0 r\big) r^{g_0/2-1} \\
&\qquad \times \prod_{i=1}^{t} \big\{ \exp\big(-\tfrac{1}{2} a_i w_i\big) w_i^{g_i/2-1} \big\}.
\end{aligned} \tag{3.1}
$$

Then the required conditional distributions are given by

$$
\mathbf{B}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{v}, r, w_1, \ldots, w_t \sim N\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{Z}\mathbf{v}), r^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\Big], \tag{3.2}
$$

$$
\begin{aligned}
\mathbf{v}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, r, w_1, \ldots, w_t \sim N\Bigg[&\Bigg(\mathbf{Z}^T\mathbf{Z} + \bigoplus_{l=1}^{t} r^{-1} w_l \mathbf{I}_{q_l}\Bigg)^{-1} \mathbf{Z}^T(\mathbf{y} - \mathbf{X}\mathbf{b}), \\
&r^{-1}\Bigg(\mathbf{Z}^T\mathbf{Z} + \bigoplus_{l=1}^{t} r^{-1} w_l \mathbf{I}_{q_l}\Bigg)^{-1}\Bigg],
\end{aligned} \tag{3.3}
$$

$$
\begin{aligned}
&R|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, w_1, \ldots, w_t \\
&\quad \sim \text{Gamma}\Big(\tfrac{1}{2}\big\{\|\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{v}\|^2 + a_0\big\}, \tfrac{1}{2}(N + g_0)\Big),
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
&W_i|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, r, w_j, j \neq i \\
&\quad \sim \text{Gamma}\Big(\tfrac{1}{2}\big(\|\mathbf{v}_i\|^2 + a_i\big), \tfrac{1}{2}(q_i + g_i)\Big), \qquad i = 1, \ldots, t,
\end{aligned} \tag{3.5}
$$

$$
\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}, \mathbf{b}, \mathbf{v}, r, w_1, \ldots, w_t \sim N\big(\mathbf{X}^{(2)}\mathbf{b} + \mathbf{Z}^{(2)}\mathbf{v}, r^{-1}\mathbf{I}_{N-n}\big). \tag{3.6}
$$

Gelfand and Smith (1990) have pointed out that it suffices to know (3.2)–(3.6) to find the joint distribution of $\mathbf{Y}^{(2)}, \mathbf{B}, \mathbf{v}, R, W_1, \ldots, W_t$ conditional on $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. Also, they have provided the recipe of finding the Monte Carlo approximation to the posterior pdf of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ on the basis of these conditional distributions. However, the procedure requires $p + q + 1 + t + N - n$ random variate generations to complete a cycle. If we run m sequences out to the ith iteration, a total of $mi\,(p + q + 1 + t + N - n)$ random variate generations are needed, and we need a great deal of total computing time. The substitution algorithm of Tanner and Wong (1987) requires even $(p + q + 1 + t + N - n)\,(p + q + t + N - n)$ random variate generations to complete a cycle in as much as other conditional distributions involving subsets of the random variables given in (3.2)–(3.6) are needed. Clearly, if the dimension of λ

is small, it is much simpler to execute direct numerical integration using one of the available packages. To carry out direct numerical integration, we have written our programs in the FORTRAN language, and have used the IMSL version 9.2 subroutine packages. A microvax computer was available for execution of our programs.

4. Data analysis. We now turn to the actual data analysis. The first set of data relates to the quality of radiation therapy care for cancer patients, while the second set of data relates to the prediction of areas under corn and soybeans for 12 counties in North Central Iowa.

4.1. *Radiation therapy data.* The data were collected with the primary objective of comparing the quality of radiation therapy for cancer patients among subpopulations of a population of facilities where radiation therapy was practiced. We have, however, used the data primarily for the comparison of several estimators of the finite population mean when two-stage sampling is performed. Our finite population of units is actually the sample units arising from a 1978 survey of patients suffering from cervical cancer. For conducting this survey, radiation therapy facilities were grouped into several strata that were thought to be relatively homogeneous in the quality of care that patients received. The five strata considered in this paper correspond to strata 1, 2, 4, 5 and 6 of Calvin and Sedransk (1991) who have provided a more detailed description of what these strata actually are. The number of facilities contained in these five strata are 10, 15, 11, 30 and 11, respectively, and are treated as primary sampling units (PSUs). Among these PSUs, we have selected a $\frac{1}{3}$ simple random sample resulting in the selection of 3, 5, 4, 10 and 4 PSUs from the five strata. From each selected PSU, with p patient records, a simple random sample of size $[\frac{1}{2}(p+1)]$ is selected, where $[u]$ denotes the integer part of u.

The present analysis considers "pretreatment" scores for each patient. For a given patient, for each disease site, a committee of experts identified a set of services and procedures (S/P's) that were thought to be of prime importance for a complete pretreatment evaluation and for planning and monitoring therapy. The committee also assigned weights (0.5 to 4.0) to these S/P's to indicate their relative importance. Then, for each patient, a score is defined by $\Sigma_i W_i^* Z_i / \Sigma_i W_i^*$, where $Z_i = 1$ if the ith S/P is performed, while $Z_i = 0$ otherwise; W_i^* is the corresponding weight. The larger the score, the closer the patient's care conforms to acceptable standards of care.

Let Y_{ijk} denote the score for the kth patient in the jth facility within the ith stratum. Although the Y_{ijk}'s lie between 0 and 1, these are weighted averages of independent Bernoulli variables, and a normal approximation due to the CLT is not totally out of the way.

We assume the model given in (2.5) with $\mathbf{b} = \mu$, the general effect, and $\mathbf{x}_{ij} = 1$. As described in Section 2, from the ith stratum, a sample of l_i $(< L_i)$ primary units is taken, while for the jth selected primary unit within the ith stratum, a sample of n_{ij} $(< N_{ij})$ subunits are selected. We denote the sample

observations by Y_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, l_i$, $i = 1, \ldots, 5$. Also, let $\mathbf{y}^{(1)}$ be the vector of sample observations, $\bar{y}_{ij} = n_{ij}^{-1}\sum_{k=1}^{n_{ij}} y_{ijk}$, $B_{ij} = \lambda_2/(\lambda_2 + n_{ij})$, $\bar{y}_i = \sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_{ij}/\sum_{j=1}^{l_i}(1 - B_{ij})$, $\alpha_i = \lambda_1/(\lambda_1 + \lambda_2\sum_{j=1}^{l_i}(1 - B_{ij}))$, $\bar{y} = \sum_{i=1}^{5}(1 - \alpha_i)\bar{y}_i/\sum_{i=1}^{5}(1 - \alpha_i)$, $f_{ij} = (N_{ij} - n_{ij})/N_{ij}$. Then the HB predictor of $\gamma_i = \sum_{j=1}^{L_i}\sum_{k=1}^{N_{ij}} Y_{ijk}/\sum_{j=1}^{L_i} N_{ij}$, the population mean for the ith stratum, is given by

$$
\begin{aligned}
e_{\mathrm{HB}}^{i} = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1} E\Bigg[& \sum_{j=1}^{l_i} N_{ij}(1 - f_{ij}B_{ij})\bar{y}_{ij} \\
& + \left\{\left(\sum_{j=l_i+1}^{L_i} N_{ij}\right) + \sum_{j=1}^{l_i} N_{ij} f_{ij} B_{ij}\right\} \\
& \times \{(1 - \alpha_i)\bar{y}_i + \alpha_i \bar{y}\} \Big| \mathbf{y}^{(1)} \Bigg].
\end{aligned} \tag{4.1.1}
$$

The posterior pdf of Λ given in (2.11) simplifies in this case to

$$
\begin{aligned}
f(\lambda_1, \lambda_2 | \mathbf{y}^{(1)}) \propto & \left(\prod_{i=1}^{m}\prod_{j=1}^{l_i} B_{ij}^{1/2}\right)\left(\prod_{i=1}^{m}\alpha_i^{1/2}\right)\left(\lambda_1 \sum_{i=1}^{m}(1 - \alpha_i)\right)^{-1/2} \\
& \times \Bigg(s + a_0 + a_1\lambda_1 + a_2\lambda_2 + \sum_{i=1}^{m} K_{3i} \\
& - \left(\sum_{i=1}^{m} K_{2i}\right)^2 \Big/ \left(\sum_{i=1}^{m} K_{1i}\right) \Bigg)^{-(n_{\cdot\cdot} + g_0 + g_1 + g_2 - 1)/2},
\end{aligned} \tag{4.1.2}
$$

where $m = 5$, $s = \sum_{i=1}^{m}\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^2$, $K_{1i} = \lambda_1(1 - \alpha_i)$, $K_{2i} = \lambda_1(1 - \alpha_i)\bar{y}_i$, $K_{3i} = \lambda_2[\sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_{ij}^2 - (1 - \alpha_i)\sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_i^2]$ and $n_{\cdot\cdot} = \sum_{i=1}^{m}\sum_{j=1}^{l_i} n_{ij}$. In finding the HB predictor, we have used (4.1.2) with $a_0 = g_0 = g_1 = g_2 = 0$, $a_1 = a_2 = 0.0005$, and have carried out two-dimensional numerical integration.

An alternative estimator of γ_i is due to Ghosh and Lahiri (1988) which uses estimates of B_{ij}'s and α_i's rather than assigning any prior distribution on R and Λ. The resulting EB estimate of γ_i is given by

$$
\begin{aligned}
e_{\mathrm{EB}}^{i} = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1} \Bigg[& \sum_{j=1}^{l_i} N_{ij}\left(1 - f_{ij}\hat{B}_{ij}\right)\bar{y}_{ij} \\
& + \left\{\sum_{j=l_i+1}^{L_i} N_{ij} + \sum_{j=1}^{l_i} N_{ij} f_{ij}\hat{B}_{ij}\right\} \\
& \times \{(1 - \hat{\alpha}_i)\bar{y}_{i*} + \hat{\alpha}_i \bar{y}_*\} \Bigg],
\end{aligned} \tag{4.1.3}
$$

where $\hat{B}_{ij} = (1 + \hat{\lambda}_2^{-1} n_{ij})^{-1}$, $\hat{\alpha}_i = \hat{\lambda}_2^{-1}(\hat{\lambda}_2^{-1} + \hat{\lambda}_1^{-1}\sum_{j=1}^{l_i}(1 - \hat{B}_{ij}))^{-1}$, $\bar{y}_{i*} = \sum_{j=1}^{l_i}(1 - \hat{B}_{ij})\bar{y}_{ij} / \sum_{j=1}^{l_i}(1 - \hat{B}_{ij})$ if $\hat{\lambda}_2^{-1} \neq 0$ and $\bar{y}_{i*} = l_i^{-1}\sum_{j=1}^{l_i}\bar{y}_{ij}$, otherwise. Similarly, $\bar{y}_* = \sum_{i=1}^{m}(1 - \hat{\alpha}_i)\bar{y}_i / \sum_{i=1}^{m}(1 - \hat{\alpha}_i)$ if $\hat{\lambda}_1^{-1} \neq 0$ and $\bar{y}_* = m^{-1}\sum_{i=1}^{m}\bar{y}_i$ otherwise. The estimators $\hat{\lambda}_1^{-1}$ and $\hat{\lambda}_2^{-1}$ are given by Ghosh and Lahiri (1988), pages 205–206.

Four other estimates of γ_i are given below. These are:

$$(4.1.4)\quad e_U^i = \left(\frac{L_i}{l_i}\right)\left(\sum_{j=1}^{l_i} N_{ij}\bar{y}_{ij}\right)\Big/\left(\sum_{j=1}^{L_i} N_{ij}\right) \qquad \text{(a design-unbiased estimate)},$$

$$(4.1.5)\quad e_R^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \sum_{j=1}^{l_i}(N_{ij} - n_{ij})\bar{y}_{ij} + \left(\sum_{j=1}^{l_i} N_{ij}\bar{y}_{ij}\Big/\sum_{j=1}^{l_i} N_{ij}\right)\left(\sum_{j=l_i+1}^{L_i} N_{ij}\right)\right]$$

(the ratio-type estimate),

$$(4.1.6)\quad e_0^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \left(\sum_{j=1}^{L_i} N_{ij} - \sum_{j=1}^{l_i} n_{ij}\right)\left(\sum_{j=1}^{l_i} n_{ij}\bar{y}_{ij}\right)\Big/\sum_{j=1}^{l_i} n_{ij}\right]$$

(the expansion estimate),

$$(4.1.7)\quad e_{\mathrm{RO}}^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \sum_{j=1}^{l_i}(N_{ij} - n_{ij})\bar{y}_{ij} + \left(\sum_{j=1}^{l_i}\bar{y}_{ij}/l_i\right)\left(\sum_{j=l+1}^{L_i} N_{ij}\right)\right] \quad \text{(Royall's estimate)}.$$

The estimates e_R, e_0 and e_{RO} are all based on predicted values of the unobserved units on the basis of the sampled units. However, in contrast to the present model, they can possibly be justified on the basis of some other models as given for example in Royall (1976). Table 1 provides the true population means as well as the six different estimates for each stratum.

The average absolute biases of the HB estimate, the EB estimate, the design unbiased estimate, the ratio-type estimate, the expansion estimate and Royall's estimate for the given data set are given respectively by 0.03102, 0.03156, 0.12932, 0.06277, 0.06009 and 0.04844. Thus the HB estimate has a slight edge over the EB estimate and much greater edge over the others in terms of average absolute bias. Also, the total sum of squared deviations of the HB estimates from the true means is 0.0085. The corresponding figures for e_{EB},

TABLE 1
The true means γ_i's and the estimates

i	γ_i	e^i_{HB}	e^i_{EB}	e^i_U	e^i_R	e^i_0	e^i_{RO}
1	0.73326	0.79789	0.80314	0.71201	0.91849	0.92190	0.93210
2	0.76149	0.76357	0.76442	0.91002	0.77214	0.76815	0.75043
3	0.74482	0.76778	0.76844	0.78208	0.78382	0.78299	0.75043
4	0.68933	0.75057	0.74971	0.89651	0.73864	0.74003	0.71533
5	0.74549	0.74130	0.74181	0.98056	0.71313	0.72653	0.71998

e_U, e_R, e_0 and e_{RO} turn out to be 0.0091, 0.1211, 0.0391, 0.0400 and 0.0409. Thus the percentage reduction in the total sum of squared deviations for the HB estimates is 6.6 in comparison with the EB estimates, 93.0 in comparison with the design unbiased estimates, 78.3 in comparison with the ratio-type estimates, 78.8 in comparison with the expansion estimates and 79.3 in comparison with Royall's estimates. An EB point estimator is usually on par with the corresponding HB point estimator. So the small improvement of the HB estimator over the EB estimator in reducing the total sum of squared deviations is not so surprising. However, the improvement of the HB estimator over the other four estimators is indeed startling. One possible explanation for this fact is that many of the other estimators are optimal under models which do not take into account variation in the primary sampling units. Our model accounts for this extra source of variation in producing more reliable estimates.

We also mention in passing that the posterior s.d.'s associated with the HB estimates in the five strata are given respectively by 0.050, 0.036, 0.043, 0.030 and 0.039.

4.2. *Prediction of areas under corn and soybeans.* Next, we analyze a data set where the objective is to predict areas under corn and soybeans for 12 counties in North Central Iowa based on the 1978 June Enumerative Survey as well as *LANDSAT* satellite data. The data set appears in Battese, Harter and Fuller (1988) who conducted a variance components analysis for this problem. The background of this problem is as follows.

The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in 37 sample segments (each segment was about 250 hectares) of 12 counties in North Central Iowa by interviewing farm operators. Based on *LANDSAT* readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels (a term for "picture element" about 0.45 hectares) in the 12 counties. The number of segments in each country, the number of hectares of corn and soybeans (as reported in the June Enumerative Survey), the number of pixels classified as corn and soybeans for each sample segment and the county mean number of pixels classified as corn and soybeans (the total number of pixels classified as that crop divided by the number of segments in that county) are reported in

Table 1 of Battese, Harter and Fuller (1988). In order to make our results comparable to that of Battese, Harter and Fuller (1988), the second segment in Hardin County was ignored.

Battese, Harter and Fuller (1988) considered the model

$$Y_{ij} = b_0 + b_1 x_{1ij} + b_2 x_{2ij} + v_i + e_{ij}, \tag{4.2.1}$$

where i is a subscript for the county and j is a subscript for a segment within the given county ($j = 1, \ldots, N_i$, the number of segments in the ith county, $i = 1, \ldots, 12$). Here Y_{ij} is the reported number of hectares of soybeans and x_{1ij} (x_{2ij}) is the number of pixels classified as corn (soybeans) for the jth segment in the ith county. They assumed (in our notation) $E(v_i) = E(e_{ij}) = 0$, $V(v_i) = (\lambda r)^{-1}$, $V(e_{ij}) = r^{-1}$, $\operatorname{cov}(v_i, e_{ij}) = 0$, $\operatorname{cov}(v_i, v_{i'}) = 0$, $i \neq i'$, $\operatorname{cov}(e_{ij}, e_{i'j'}) = 0$ if $(i, j) \neq (i', j')$. First, assuming λ and r known, these authors obtained BLUPs of $\mu_i = b_0 + b_1 \bar{x}_{1i(p)} + b_2 \bar{x}_{2i(p)} + v_i$, $i = 1, \ldots, 12$, where $\bar{x}_{\alpha i(p)} = N_i^{-1} \sum_{j=1}^{N_i} x_{\alpha ij}$, $\alpha = 1, 2$. Then, using Henderson's method III, they obtained estimates of the variance components, and their final predictors involved the estimated variance components. [For details, see Battese, Harter and Fuller (1988).] Henderson's method being an ANOVA method could lead to negative estimates of λ^{-1}. If this were the case, Battese, Harter and Fuller set it equal to 0. This phenomenon is likely to happen, particularly when the number of small areas or strata is small.

In this particular example, we have $t = 1$, $\lambda_1 = \lambda$, $\mathbf{D}(\boldsymbol{\lambda}) = \lambda^{-1}\mathbf{I}_m$, $\boldsymbol{\Psi} = \mathbf{I}_N$. Then $\boldsymbol{\Sigma}_{11} = \operatorname{Diag}(\mathbf{I}_{n_1} + \lambda^{-1}\mathbf{J}_{n_1}, \ldots, \mathbf{I}_{n_m} + \lambda^{-1}\mathbf{J}_{n_m})$ so that $|\boldsymbol{\Sigma}_{11}| = \prod_{i=1}^{m}\{(\lambda + n_i)/\lambda\}$. Also, writing $\bar{\mathbf{x}}_i = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$, $i = 1, \ldots, m$, one gets

$$\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sum_{i=1}^{m} n_i^2(n_i + \lambda)^{-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T \tag{4.2.2}$$

$$= \mathbf{H}(\lambda) \quad \text{(say)}.$$

Next, writing $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$, one gets

$$\mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \lambda\sum_{i=1}^{m} n_i(n_i + \lambda)^{-1}\bar{y}_i^2$$

$$- \left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(y_{ij} - n_i(n_i + \lambda)^{-1}\bar{y}_i\right)\right\}^T \mathbf{H}^{-1}(\lambda) \tag{4.2.3}$$

$$\times\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(y_{ij} - n_i(n_i + \lambda)^{-1}\bar{y}_i\right)\right\}$$

$$= Q_0(\lambda) \quad \text{(say)}.$$

The conditional pdf $f(\lambda|\mathbf{y}^{(1)})$ given in (2.11) simplifies to

$$f(\lambda|\mathbf{y}^{(1)}) \propto \lambda^{(m+g_1)/2-1} \prod_{i=1}^{m} (\lambda + n_i)^{-1/2} |\mathbf{H}(\lambda)|^{-1/2} \times (a_0 + a_1\lambda + Q_0(\lambda))^{-(n+g_0+g_1-p)/2}. \tag{4.2.4}$$

The posterior means and variances of the finite population means are now obtained from (2.8)–(2.10), (2.12)–(2.13), (4.2.2)–(4.2.4) and using the formulas for iterated conditional expectations and variances.

REMARK 1. Let $V_1(\mathbf{y}^{(1)})$ and $V_2(\mathbf{y}^{(1)})$ denote respectively the variance of the conditional expectation and expectation of the conditional variance of the finite population mean. A naive empirical Bayes procedure effectively ignores V_1 and can lead to serious underestimate of the variance. A HB procedure on the other hand rectifies this deficiency. Battese, Harter and Fuller have a frequentist approach which also incorporates the uncertainty of estimating the variance components into account.

We find the posterior means and variances of the population means for the 12 counties. Our approach eliminates the possibility of obtaining zero estimates of the variance components. The improper prior with $a_0 = a_1 = 0.005$, $g_0 = g_1 = 0$ is used for predicting areas under soybeans.

Table 2 provides the HB predictors (e_{HB}), the EB predictors (e_{EB}), the BHF predictors (e_{BHF}) and the associated standard errors s_{HB}, s_{EB} and s_{BHF}, respectively. Note that the EB predictors are obtained by replacing λ with its Henderson's Method III estimate in $E[N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}|\mathbf{y}^{(1)}, \lambda]$. Also, we provide the V_1 and V_2 values to demonstrate that V_1 can sometimes contribute significantly toward the posterior variance.

As one might anticipate, e_{HB} and e_{EB} are extremely close as point predictors; e_{BHF} differs from e_{EB} because it uses a different estimate of λ, and

TABLE 2
The predicted hectares of soybeans and standard errors

County	e_{HB}	e_{EB}	e_{BHF}	s_{HB}	s_{EB}	s_{BHF}	V_1	V_2
Cerro Gordo	78.8	78.2	77.5	11.7	11.6	12.7	7.67	128.59
Franklin	67.1	65.9	64.8	8.2	7.5	7.8	11.94	54.92
Hamilton	94.4	94.6	95.0	11.2	11.4	12.4	1.97	123.61
Hancock	100.4	100.8	101.1	6.2	6.1	6.3	1.35	37.59
Hardin	75.4	75.1	74.9	6.5	6.4	6.6	0.37	41.84
Humboldt	81.9	80.6	79.2	10.4	9.3	10.0	22.62	85.40
Kossuth	118.2	119.2	120.2	6.6	6.0	6.2	7.99	36.23
Pocahontas	113.9	113.7	113.8	7.5	7.5	7.9	0.06	55.98
Webster	110.0	109.7	109.6	6.6	6.6	6.8	0.64	43.91
Winnebago	97.3	98.0	98.7	7.7	7.5	7.9	4.11	55.70
Worth	87.8	87.2	86.6	11.1	11.1	12.1	4.06	118.17
Wright	111.9	112.4	112.9	7.7	7.6	8.0	1.62	57.48

thereby leads to slightly different predicted values. It is important to note that the difference between e_{BHF} and either e_{EB} or e_{HB} is much more pronounced than any difference between e_{HB} and e_{EB}.

The naive EB estimator, in general, underestimates the standard error in comparison with the HB estimator. With the exception of Hamilton County, s_{EB} is always smaller or equal to s_{HB}. The difference can be significant as evidenced from the figures given in Humboldt County where s_{EB} is about 10% smaller than s_{HB}.

However, s_{HB} and s_{BHF} are both very good as estimates of standard errors. In this example, while s_{BHF} is never smaller than s_{HB} by more than 6.1%, it can exceed s_{HB} by about 9.7%.

One may wonder whether the proposed HB predictors which perform so well conditionally enjoy any frequentist properties. To answer this, we undertook an extensive simulation study using the BHF model. The detailed results are not reported in this paper, but our findings indicated that the simulated mean squared errors for the HB predictors were matching those for the BHF predictors up to the fifth decimal place, while (1.96) s.d. coverage probabilities turned out to be slightly bigger for HB than BHF, both being very close to 95% under all circumstances.

5. The HB predictor in a special case. We consider in this section the special case when $\boldsymbol{\lambda}$ is known, while $\mathbf{B}$ and R are independently distributed with $\mathbf{B} \sim \text{uniform}(R^p)$ and $R \sim \text{Gamma}(\frac{1}{2}a_0, \frac{1}{2}g_0)$. We are still interested in finding the posterior distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. Recall the notation $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{G}$ given in (2.8)–(2.10). Since $\boldsymbol{\lambda}$ is known in this case, we have the following Theorem 2 instead of Theorem 1.

THEOREM 2. *Assume that $n + g_0 > p + 2$. Then under the model given in* (A) *and* (B) *with $\boldsymbol{\lambda}$ known, and an independent uniform (R^p) prior for $\mathbf{B}$ and a* Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$ *prior for R, the conditional distribution of $Y^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ is multivariate-t with location parameter $\mathbf{M}\mathbf{y}^{(1)}$, scale parameter $(n + g_0 - p)^{-1}(a_0 + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)})\mathbf{G}$ and degrees of freedom $n + g_0 - p$.*

The proof of Theorem 2 is similar to the proof of the first part of Theorem 1 provided in the Appendix and is omitted. Using the properties of the multivariate-t distribution, it is now possible to obtain closed-form expressions for $E[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}]$ and $V[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}]$, where $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \mathbf{A}\mathbf{Y}^{(1)} + \mathbf{C}\mathbf{Y}^{(2)}$. In particular, the Bayes estimate of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ under any quadratic loss is now

$$\mathbf{e}_B^*(\mathbf{y}^{(1)}) = (\mathbf{A} + \mathbf{C}\mathbf{M})\mathbf{y}^{(1)}. \tag{5.1}$$

We may note that the posterior mean given in (5.1) does not depend on the prior distribution of R.

There are alternative ways to generate the same predictor $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Suppose, for example, one assumes only (2.1) or (2.2) with $\mathbf{b}$ known (r may or may not be known). Then the best predictor (best linear

predictor without the normality assumption) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ in the sense of having the smallest mean squared error matrix is given by

$$E_\theta\left[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\middle|\mathbf{Y}^{(1)}\right] = \mathbf{C}\left[\Sigma_{21}\Sigma_{11}^{-1}\mathbf{Y}^{(1)} + \left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\mathbf{b}\right] + \mathbf{A}\mathbf{Y}^{(1)} \quad (\text{a.e. } \mathbf{Y}^{(1)}), \tag{5.2}$$

where $\theta = (\mathbf{b}^T, r)^T$.

[We say that $\mathbf{E} \le \mathbf{F}$ for two symmetric matrices $\mathbf{E}$ and $\mathbf{F}$ if $\mathbf{F} - \mathbf{E}$ is nonnegative definite (n.n.d.).] If $\mathbf{b}$ is unknown, then one replaces $\mathbf{b}$ by its UMVUE (BLUE without the normality assumption)

$$\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{Y}^{(1)}.$$

The resulting predictor of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ turns out to be $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$. In this sense, $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is also an empirical Bayes predictor of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Harville (1985, 1988, 1990) recognized this for predicting scalars.

We shall now discuss some frequentist properties of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$. First, we assume the normal model (2.1) or (2.2) with λ known. No prior distribution for $\mathbf{B}$ and R is assumed, and $\theta = (\mathbf{b}^T, r)^T$ is treated as an unknown parameter. We prove the optimality of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ within the class of all unbiased predictors of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. This result is then used to prove the optimality of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ once again within the class of all unbiased predictors of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ for a class of spherically symmetric distributions of $\mathbf{Y}$ including but not limited to the normal distribution.

We start with the following definition.

DEFINITION 1. A predictor $\mathbf{T}(\mathbf{Y}^{(1)})$ is said to be a *best unbiased predictor* (BUP) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ if $E_\theta[\mathbf{T}(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ and for every predictor $\delta(\mathbf{Y}^{(1)})$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ satisfying $E_\theta[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ, $V_\theta[\mathbf{T}(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] \le V_\theta[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})]$ for all θ provided the quantities are finite.

The following theorem is proved.

THEOREM 3. *Under the model* (2.1) *or* (2.2), $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ *is the BUP of* $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

PROOF. Write $\mathbf{H}_0 = \mathbf{A} + \mathbf{C}\Sigma_{21}\Sigma_{11}^{-1}$ and $\mathbf{U} = \mathbf{C}[\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}]$. Then, from (5.2), $E[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)}] = \mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}$ a.e. $(\mathbf{Y}^{(1)})$. For an arbitrary predictor $\delta(\mathbf{Y}^{(1)})$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$, write

$$\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \left[\delta(\mathbf{Y}^{(1)}) - \left(H_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}\right)\right] + \left[\left(\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}\right) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\right]. \tag{5.3}$$

Then, from (5.2) and (5.3),

$$
\begin{aligned}
&E_\theta\Big[\{\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\Big] \\
&\quad = E_\theta\Big[\{(\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)}) - \mathbf{U}\mathbf{b}\} \\
&\qquad \times\{(\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)}) - \mathbf{U}\mathbf{b}\}^T\Big] \qquad (5.4) \\
&\qquad + E_\theta\Big[\{\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\} \\
&\qquad\quad \times\{\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\Big].
\end{aligned}
$$

Hence minimization of the left-hand side of (5.4) wrt $\delta(\mathbf{Y}^{(1)})$ amounts to the minimization of the first term in the right-hand side of (5.4) wrt $\delta(\mathbf{Y}^{(1)})$. Since $\mathbf{Y}^{(1)} \sim N(\mathbf{X}^{(1)}\mathbf{b}, r^{-1}\Sigma_{11})$, from the classical theory of least squares it follows that the first term in the right-hand side of (5.4) is minimized wrt $\delta(\mathbf{Y}^{(1)})$ if and only if $\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)} = \mathbf{U}(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{Y}^{(1)}$ a.e. $(\mathbf{Y}^{(1)})$, that is, $\delta(\mathbf{Y}^{(1)}) = (\mathbf{A} + \mathbf{CM})\mathbf{Y}^{(1)} = \mathbf{e}_B^*(\mathbf{Y}^{(1)})$ a.e. $(\mathbf{Y}^{(1)})$. The proof of Theorem 3 is complete. □

REMARK 2. It follows from the proof of the theorem that the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ is unique with probability 1.

REMARK 3. It is possible to generalize Theorem 2 for a more general class of distributions of $\mathbf{Y}$. Suppose that conditional on $R = r$, $\mathbf{Y} \sim N(\mathbf{X}\mathbf{b}, r^{-1}\Sigma)$, while marginally R has any proper distribution. The objective is once again to minimize the left-hand side of (5.4). We achieve this by first computing this expectation conditional on $R = r$. We may note that $E[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}, R = r] = \mathbf{H}_0\mathbf{y}^{(1)} + \mathbf{U}\mathbf{b}$ *does not* depend on r. Hence we obtain an identity similar to (5.4) conditional on $R = r$, and as in the proof of Theorem 3, conclude that $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

Next we dispense with any distributional assumption in (2.1) and show that $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ within the class of all linear unbiased predictors. A predictor $\delta(\mathbf{Y}^{(1)})$ is said to be linear if $\delta(\mathbf{Y}^{(1)})$ has the form $\mathbf{H}\mathbf{Y}^{(1)}$ for some known $u \times n$ matrix $\mathbf{H}$. If, in addition, $E_\theta[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ, we say that $\delta(\mathbf{Y}^{(1)})$ is a *linear unbiased predictor* (LUP) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. We now introduce another definition.

DEFINITION 2. A LUP $\mathbf{P}\mathbf{Y}^{(1)}$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ is said to be a *best linear unbiased predictor* (BLUP) if for every LUP $\mathbf{H}\mathbf{Y}^{(1)}$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$, $V_\theta(\mathbf{H}\mathbf{Y}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})) - V_\theta(\mathbf{P}\mathbf{Y}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}))$ is n.n.d. for all θ.

We now prove the following theorem.

THEOREM 4. *Consider the model* (2.2) *and assume that* $E_\theta[\mathbf{e}] = \mathbf{0}$, $E_\theta[\mathbf{v}] = \mathbf{0}$, $E_\theta[\mathbf{e}v^T] = \mathbf{0}$, $E_\theta[\mathbf{e}^T\mathbf{e}] < \infty$ *and* $E_\theta[\mathbf{v}^T\mathbf{v}] < \infty$. *Then* $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ *is the BLUP of* $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

PROOF. Suppose $\mathbf{WY}^{(1)}$ is an unbiased predictor of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Then $\mathrm{E}_\theta[\mathbf{WY}^{(1)} - (A\mathbf{Y}^{(1)} + C\mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ, which is equivalent to $(\mathbf{W} - \mathbf{A})\mathbf{X}^{(1)} = \mathbf{CX}^{(2)} = \mathbf{CMX}^{(1)}$ from (2.8) and (2.9), that is, $(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{X}^{(1)} = \mathbf{0}$. Next write

$$\begin{aligned}\mathbf{WY}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) &= \mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)}) + \mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) \\ &= (\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{Y}^{(1)} + \mathbf{C}(\mathbf{MY}^{(1)} - \mathbf{Y}^{(2)}).\end{aligned} \tag{5.5}$$

Observe next that since $\mathbf{MX}^{(1)} = \mathbf{X}^{(2)}$,

$$\begin{aligned}&E_\theta\left[\mathbf{C}(\mathbf{MY}^{(1)} - \mathbf{Y}^{(2)})\{(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{Y}^{(1)}\}^T\right] \\ &\quad = E_\theta\left[\mathbf{C}\{\mathbf{M}(\mathbf{Y}^{(1)} - E_\theta(\mathbf{Y}^{(1)})) - (\mathbf{Y}^{(2)} - E_\theta(\mathbf{Y}^{(2)}))\}\right. \\ &\qquad\qquad \left.\times\mathbf{Y}^{(1)T}(\mathbf{W} - \mathbf{A} - \mathbf{CM})^T\right] \\ &\quad = E_\theta\left[\mathbf{C}(\mathbf{M}\Sigma_{11} - \Sigma_{21})(\mathbf{W} - \mathbf{A} - C\mathbf{M})^T\right].\end{aligned} \tag{5.6}$$

But, using (2.8) and (2.9),

$$\mathbf{M}\Sigma_{11} - \Sigma_{21} = (\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)})(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}. \tag{5.7}$$

Since $\mathbf{X}^{(1)T}(\mathbf{W} - \mathbf{A} - \mathbf{CM})^T = [(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{X}^{(1)}]^T = \mathbf{0}$, it follows from (5.6) and (5.7) that the left-hand side of (5.6) is $\mathbf{0}$. Now, from (5.5),

$$\begin{aligned}&E_\theta\left[\{\mathbf{WY}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{WY}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\right] \\ &\quad = E_\theta\left[\{\mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)})\}\{\mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)})\}^T\right] \\ &\qquad + E_\theta\left[\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\right] \\ &\quad \geq E_\theta\left[\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\right]\end{aligned}$$

with equality if and only if $\mathbf{WY}^{(1)} = \mathbf{e}_B^*(\mathbf{Y}^{(1)})$ a.e. $(\mathbf{Y}^{(1)})$. The proof of Theorem 4 is complete. □

APPENDIX

PROOF OF THEOREM 1. Under the assumptions of the theorem, the joint pdf of $\mathbf{Y}$, $\mathbf{B}$, R and Λ is given by

$$\begin{aligned}&f(\mathbf{y}, \mathbf{b}, r, \boldsymbol{\lambda}) \\ &\quad \propto r^{N/2}|\Sigma|^{-1/2}\exp\left[-\tfrac{1}{2}r(\mathbf{y} - \mathbf{Xb})^T\Sigma^{-1}(\mathbf{y} - \mathbf{Xb})\right]\exp(-\tfrac{1}{2}a_0 r)r^{g_0/2-1} \\ &\qquad \times \exp\left(-\tfrac{1}{2}r\sum_{i=1}^{t} a_i\lambda_i\right)\prod_{i=1}^{t}(\lambda_i r)^{g_i/2-1}r^t \\ &\quad = |\Sigma|^{-1/2}\exp\left[-\tfrac{1}{2}r\left\{(\mathbf{y} - \mathbf{Xb})^T\Sigma^{-1}(\mathbf{y} - \mathbf{Xb}) + a_0 + \sum_{i=1}^{t} a_i\lambda_i\right\}\right] \\ &\qquad \times r^{(N+\Sigma_{i=0}^{t}g_i)/2-1}\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.\end{aligned} \tag{A.1}$$

Now

$$
\begin{aligned}
&(\mathbf{y}-\mathbf{X}\mathbf{b})^T\Sigma^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}) \\
&\quad=\left[\mathbf{b}-(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}\right]^T(\mathbf{X}^T\Sigma^{-1}\mathbf{X}) \\
&\qquad\times\left[\mathbf{b}-(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}\right]+\mathbf{y}^T\mathbf{Q}\mathbf{y},
\end{aligned}
\tag{A.2}
$$

where $\mathbf{Q}=\Sigma^{-1}-\Sigma^{-1}\mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}$. From (A.1) and (A.2), one gets the joint pdf of $\mathbf{Y}$, R and Λ given by

$$
\begin{aligned}
f(\mathbf{y},r,\boldsymbol{\lambda}) \propto\ &|\Sigma|^{-1/2}|\mathbf{X}^T\Sigma^{-1}\mathbf{X}|^{-1/2}r^{(N+\Sigma_{i=0}^{t}g_i-p)2-1} \\
&\times\exp\left[-\tfrac{1}{2}r\left(a_0+\Sigma_{i=1}^{t}a_i\lambda_i+\mathbf{y}^T\mathbf{Q}\mathbf{y}\right)\right]\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.
\end{aligned}
\tag{A.3}
$$

Now, integrating wrt R, one finds the pdf of $\mathbf{Y}$ and Λ given by

$$
\begin{aligned}
f(\mathbf{y},\boldsymbol{\lambda}) \propto\ &|\Sigma|^{-1/2}|\mathbf{X}^T\Sigma^{-1}\mathbf{X}|^{-1/2}\left(a_0+\sum_{i=1}^{t}a_i\lambda_i+\mathbf{y}^T\mathbf{Q}\mathbf{y}\right)^{-(N+\Sigma_{i=0}^{t}g_i-p)/2} \\
&\times\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.
\end{aligned}
\tag{A.4}
$$

Now, using a standard formula for partitioned matrices [e.g., Searle (1971), page 46], we have

$$
\mathbf{y}^T\Sigma^{-1}\mathbf{y}=\mathbf{y}^{(1)T}\Sigma_{11}^{-1}\mathbf{y}^{(1)}+\left(\mathbf{y}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}\right)^T\Sigma_{22.1}^{-1}\left(\mathbf{y}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}\right).
\tag{A.5}
$$

Similarly,

$$
\begin{aligned}
\mathbf{y}^T\Sigma^{-1}\mathbf{X}&=\mathbf{y}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)} \\
&\quad+\left(\mathbf{y}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}\right)^T\Sigma_{22.1}^{-1}\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right) \\
&=\mathbf{t}_1^T+\mathbf{t}_2^T\quad\text{(say)},
\end{aligned}
\tag{A.6}
$$

$$
\begin{aligned}
\mathbf{X}^T\Sigma^{-1}\mathbf{X}&=\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)} \\
&\quad+\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^T\Sigma_{22.1}^{-1}\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right).
\end{aligned}
\tag{A.7}
$$

Using the matrix inversion formula [see Exercise 2.9, page 33 of Rao (1973)], we have from (A.7) that

$$
\begin{aligned}
&(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1} \\
&\quad=\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}-\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^T \\
&\qquad\times\Big\{\Sigma_{22.1}+\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1} \\
&\qquad\qquad\times\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{21}^{-1}\mathbf{X}^{(1)}\right)^T\Big\}^{-1} \\
&\qquad\times\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1} \\
&\quad=\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}-\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^T\mathbf{G}^{-1} \\
&\qquad\times\left(\mathbf{X}^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\quad\text{by (2.10)} \\
&\quad=\mathbf{M}_1-\mathbf{M}_2\quad\text{(say)}.
\end{aligned}
\tag{A.8}
$$

From (A.6), (A.8) and (2.8)–(2.10), we get after simplifications

$$\text{(A.9)}\quad \mathbf{y}^T\Sigma^{-1}\mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y} = \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_1 - \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_1 + \mathbf{t}_2^T\mathbf{M}_1\mathbf{t}_2 - \mathbf{t}_2^T\mathbf{M}_2\mathbf{t}_2 + 2\mathbf{t}_1^T(\mathbf{M}_1 - \mathbf{M}_2)\mathbf{t}_2,$$

$$\text{(A.10)}\quad \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_1 = \mathbf{y}^{(1)T}(\Sigma_{11}^{-1} - \mathbf{K})\mathbf{y}^{(1)},$$

$$\text{(A.11)}\quad \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_1 = (\mathbf{M}\mathbf{y}^{(1)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)})^T\mathbf{G}^{-1}(\mathbf{M}\mathbf{y}^{(1)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}),$$

$$\text{(A.12)}\quad \mathbf{t}_2^T\mathbf{M}_1\mathbf{t}_2 = (\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)})^T[\Sigma_{22.1}^{-1}\mathbf{G}\Sigma_{22.1}^{-1} - \Sigma_{22.1}^{-1}] \times (\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}),$$

$$\text{(A.13)}\quad \mathbf{t}_2^T\mathbf{M}_2\mathbf{t}_2 = (\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)})^T \times [\Sigma_{22.1}^{-1}\mathbf{G}\Sigma_{22.1}^{-1} - 2\Sigma_{22.1}^{-1} + \mathbf{G}^{-1}](\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}),$$

$$\text{(A.14)}\quad \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_2 = (\mathbf{M}\mathbf{y}^{(1)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)})^T\Sigma_{22.1}^{-1}(\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}),$$

$$\text{(A.15)}\quad \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_2 = (\mathbf{M}\mathbf{y}^{(1)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)})^T[\Sigma_{22.1}^{-1} - \mathbf{G}^{-1}] \times (\mathbf{y}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}^{(1)}),$$

Using the same definition of $\mathbf{Q}$, it follows from (A.5)–(A.15) with some algebraic manipulations that

$$\text{(A.16)}\quad \mathbf{y}^T\mathbf{Q}\mathbf{y} = \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)} + (\mathbf{y}^{(2)} - \mathbf{M}\mathbf{y}^{(1)})^T\mathbf{G}^{-1}(\mathbf{y}^{(2)} - \mathbf{M}\mathbf{y}^{(1)}).$$

Combining (A.4), (A.16) and (2.6), one gets the first part of Theorem 1.

Now to find the conditional distribution of Λ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$, one can have as in (A.4) that the pdf of $\mathbf{Y}^{(1)}$ and Λ is given by

$$\text{(A.17)}\quad f(\mathbf{y}^{(1)}, \boldsymbol{\lambda}) \propto |\Sigma_{11}|^{-1/2}|\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}|^{-1/2} \times \left(a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right)^{-(n+\sum_{i=0}^{t} g_i - p)/2} \prod_{i=1}^{t} \lambda_i^{g_i/2-1}.$$

Since $f(\boldsymbol{\lambda}|\mathbf{y}^{(1)}) \propto f(\mathbf{y}^{(1)}, \boldsymbol{\lambda})$, (2.11) follows from (A.17). □

Acknowledgments. We express our indebtedness to Professors James Calvin and Joseph Sedransk for supplying us with the set of data used for two-stage sampling. The revision has benefitted much from the very helpful comments of an Associate Editor and four referees. Thanks are due to Dr. Li-Chu Lee for carrying out some numerical computations.

REFERENCES

ALBERT, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83** 1037–1044.

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. *J. Amer. Statist. Assoc.* **83** 28–36.

CALVIN, J. and SEDRANSK, J. (1991). The patterns of care studies. *J. Amer. Statist. Assoc.* **86** 36–48.

CHOUDHRY, G. H. and RAO, J. N. K. (1988). Evaluation of small area estimations: An empirical study. Preprint.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd. ed. Wiley, New York.

DEMPSTER, A. P., RUBIN, D. B. and TSUTAKAWA, R. K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* **76** 341–353.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

GELFAND, A. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

GHOSH, M. and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* **82** 1153–1162.

GHOSH, M. and LAHIRI, P. (1988). Bayes and empirical Bayes analysis in multistage sampling. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 195–212. Springer, New York.

GHOSH, M. and LAHIRI, P. (1989). A hierarchical Bayes approach to small area estimation with auxiliary information. In *Proceedings of the Joint Indo–U.S. Workshop on Bayesian Inference in Statistics and Econometrics*. To appear.

GHOSH, M. and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.* **81** 1058–1062.

GHOSH, M. and RAO, J. N. K. (1991). Small area estimation. Technical Report 390, Dept. Statistics, Univ. Florida.

HARVILLE, D. A. (1985). Decomposition of prediction error. *J. Amer. Statist. Assoc.* **80** 132–138.

HARVILLE, D. A. (1988). Mixed-model methodology: Theoretical justifications and future directions. In *Proceedings of the Statistical Computing Section* 41–49. Amer. Statist. Assoc., Alexandria, Va.

HARVILLE, D. A. (1990). BLUP and beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock*. (D. Gianola and K. L. Hammond, eds.) 239–276. Springer, New York.

HARVILLE, D. A. and JESKE, D. R. (1989). Mean squared error of estimation and prediction under a general linear model. Preprint 89-37, Statistics Laboratory, Iowa State Univ.

HENDERSON, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (W. D. Hanson and M. F. Robinson, eds.) 141–163. Publication 982, NAS-NRC, Washington, D. C.

KACKAR, R. N. and HARVILLE, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.* **79** 853–862.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

LUI, K. J. and CUMBERLAND, W. G. (1989). A Bayesian approach to small domain estimation. *Journal of Official Statistics* **5** 143–156.

MALEC, D. and SEDRANSK, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *J. Amer. Statist. Assoc.* **80** 897–902.

MORRIS, C. N. (1988). Determining the accuracy of Bayesian empirical Bayes estimates in the familiar exponential families. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 251–263. Springer, New York.

PRASAD, N. G. N. and RAO, J. N. K. (1990). On the estimation of mean square error of small area predictors. *J. Amer. Statist. Assoc.* **85** 163–171.

PRESS, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston, New York.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

ROYALL, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.* **71** 657–664.

ROYALL, R. M. (1978). Prediction models in small area estimation. In *Synthetic Estimates for Small Areas* (J. Steinberg, ed.) 63–87. NIDA Monograph Series 24. Dept. Health, Education and Welfare, Washington, D.C.

SCOTT, A. and SMITH, T. M. F. (1969). Estimation in multistage surveys. *J. Amer. Statist. Assoc.* **64** 830–840.

SEARLE, S. R. (1971). *Linear Models*. Wiley, New York.

STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics: An International Symposium*. (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 124–137. Wiley, New York.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611

Jackknifing The Mean Squared Error of Empirical Best Predictor

Jiming Jiang, Partha Lahiri, and Shu-Mei Wan
Department of Statistics
Case Western Reserve University
Cleveland, OH 44106-7054, USA
jiang@eureka.cwru.edu

Department of Mathematics and Statistics
University of Nebraska at Lincoln
Lincoln, NE 68588-0323, USA
plahiri@mathstat.unl.edu

1. Introduction. In recent years, Empirical Best Linear Unbiased Prediction (EBLUP) approach has received considerable importance in producing small-area statistics. This is really a special case of Empirical Best Prediction (EBP) approach which can be applied even when we do not have mixed *linear* model need for EBLUP approach. The main focus of this paper is to develop a general theory for EBP approach.

We develop a suitable jackknife technique to estimate the MSE of EBP of any general mixed effect for a general model. The proposed jackknife method is very simple to implement and does not require the derivation of different derivatives needed in the Taylor series method. Thus the method should be very attractive to the practitioners. The general model we consider covers not only the mixed linear model but also many complex models like generalized linear mixed model. So long as one can get expression for EBP, our method can be applied. For example, we no longer require the assumption of normality to estimate the MSE of EBLUP given in Prasad and Rao (1990) - we just need the assumption of *posterior linearity* (see, e.g., Ghosh and Meeden 1997) which is needed anyway to justify the use of EBLUP (which is identical with linear empirical Bayes (LEB) estimator). In addition, the proposed jackknife method will work for a general M-estimator of the model parameters (which includes ML, REML and ANOVA).

The properties of the jackknife estimators have been studied extensively in the literature (see, e.g., Shao and Tu (1995)). However, the problems discussed in the paper are not currently available in the literature. First, our main interest is not in the estimation of a fixed parameter but in the prediction of a random vector which may be associated with unknown parameters. This is, of course, a more complicated problem. Secondly, even for estimating the fixed parameters, our jackknife estimator is not based on i.i.d. observations and it is not associated with the regression estimator. Furthermore, since we consider a specific class of estimators, namely, the M-estimators, the conditions under which the asymptotic results hold will be easier to verify. As will be seen, the asymptotic unbiasedness of the jackknife MSE estimator is proved essentially under some moment conditions. Thirdly, our M-estimators are more general than those considered by Reeds (1978) in the sense that ours also include the modified profile MLE (e.g., REML estimators), penalized MLE, or M-estimators not associated with a maximization process (e.g., the method of moment estimators).

Section 2 discusses the model and the proposed EBP. In section 3, we propose a jackknife method to measure the uncertainty of the proposed EBP. The asymptotic properties of our jackknife MSE estimator are also stated in this section. The mixed linear models and mixed logistic models which are important special cases of our general model are discussed in sections 2 and 4, respectively. Due to lack of space, we refer to Jiang, Lahiri and Wan (1998) for proofs of all the technical results.

2. Empirical best predictor. Let $Y_1, \ldots, Y_m$ be independent (vector-valued) observations

whose distributions depend on a vector $\phi = (\phi_k)_{1\le k\le s}$ of unknown parameters. We are interested in predicting an unobservable random vector $\theta = (\theta_l)_{1\le l\le t}$ based on $Y = (Y_j)_{1\le j\le m}$. Supopose that, when ϕ is known, the best predictor in terms of MSE is $\tilde{\theta} = E(\theta|Y) = \pi(Y_S;\phi) = (\pi_l(Y_S;\phi))_{1\le l\le t}$, where S is a subset of $\{1,\ldots,m\}$ and $Y_S = (Y_j)_{j\in S}$. For example, in small-area estimation, one is interested in predicting a mixed effect $\theta = h'\beta + \lambda' v$, where h and λ are known vectors, β is a vector of unknown fixed effects, and v is a vector of unobservable small-area specific random effects. In particular, a mixed effect associated with the ith small-area is of the form $\theta = h'\beta + v_i$. Assuming that the random effects corresponding to the small-areas are independent, the best predictor of θ is of the form $\pi(Y_i;\phi)$, where Y_i is the vector of observations associated with the ith small-area, and ϕ is the combination of β and a vector ψ of variance components.

Since ϕ is usually unknown, it is naturally replaced by an estimator, $\hat{\phi}$. The resulting predictor, $\hat{\theta} = \pi(Y_S;\hat{\phi})$ is called the empirical best predictor of θ. The estimator $\hat{\phi}$ of particular interest in this paper is an M-estimator (Huber (1981)), which is associated with a solution $\dot{\phi} = (\dot{\phi}_k)_{1\le k\le s}$ to the following equation:

$$F(\phi;Y) \;=\; \sum_{j=1}^{m} f_j(\phi;Y_j) + a(\phi) \;=\; 0\,. \tag{1}$$

In the above, $f_j(\phi;Y_j) = (f_{j,k}(\phi;Y_j))_{1\le k\le s}$ are vector-valued functions such that $Ef_j(\phi;Y_j) = 0$ when ϕ is the true parameter vector, and $a(\phi)$ is a vector-valued function which may depend on the joint distribution of $Y = (Y_j)_{1\le j\le m}$. When $a(\phi)\neq 0$, it plays the role of a modifier or penalizer.

Example 1. (ML estimator in mixed linear models) Consider a mixed linear model

$$Y_i \;=\; X_i\beta + Z_i v_i + e_i\,,\quad i = 1,\ldots,m\,, \tag{2}$$

where X_i $(n_i\times p)$ and Z_i $(n_i\times b_i)$ are known matrices, v_i and e_i are independently distributed with $v_i \overset{\text{ind}}{\sim} (0,G_i)$ and $e_i \overset{\text{ind}}{\sim} (0,R_i)$, $i = 1,\ldots,m$. Assume that $G_i = G_i(\psi)$ $(b_i\times b_i)$ and $R_i = R_i(\psi)$ $(n_i\times n_i)$ possibly depend on $\psi = (\psi_l)_{1\le l\le q}$, a $q\times 1$ vector of variance components. The ML estimator of $\phi = (\beta'\ \psi')'$ is defined as solution to the ML equations. Note that this definition does not require normality, i.e., the ML equations are used even if the data is not normal (Jiang (1996)). It is easy to show that the ML estimator of ϕ is solution to (1) where $a(\phi) = 0$, $(f_{j,k}(\phi,Y_j))_{1\le k\le p} = X_j'\Sigma_j^{-1}(\psi)(Y_j - X_j\beta)$, where $\Sigma_j(\psi) = R_j(\psi) + Z_jG_j(\psi)Z_j' = \mathrm{Var}(Y_j)$, and for $1\le l\le q$

$$f_{j,p+l}(\phi,Y_j) = (Y_j - X_j\beta)'\Sigma_j^{-1}(\psi)\left(\frac{\partial\Sigma_j}{\partial\psi_l}\right)\Sigma_j^{-1}(\psi)(Y_j - X_j\beta) - \mathrm{tr}\left(\Sigma_j^{-1}(\psi)\frac{\partial\Sigma_j}{\partial\psi_l}\right)\,.$$

Example 2. (REML estimator in mixed linear models) Similarly, the REML estimator $\dot{\psi}$ of ψ is defined as solution to the REML equations (Jiang (1996)), and the REML estimator of β as the EBLUE $\dot{\beta} = (X'\Sigma^{-1}(\dot{\psi})X)^{-1}X'\Sigma^{-1}(\dot{\psi})Y$, where $X = \mathrm{col}_{1\le i\le m}(X_i)$, $Y = \mathrm{col}_{1\le i\le m}(Y_i)$, $\Sigma(\psi) = R + ZGZ'$, $Z = \mathrm{diag}_{1\le i\le m}(Z_i)$, $G = \mathrm{diag}_{1\le i\le m}(G_i)$, and $R = \mathrm{diag}_{1\le i\le m}(R_i)$. Again, this definition does not require normality. By the identity (e.g., Searle *et al* (1992), page 451) $\Sigma^{-1} = \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} + A(A'\Sigma A)^{-1}A'$, which holds for any $N\times(N-p)$ matrix A of full rank (N is the dimension of Y) such that $A'X = 0$, it is easy to show that the REML estimator of ϕ is solution to (1) where the f_j's are the same as in Example 1; $a(\phi) = (a_k(\phi))_{1\le k\le p+q}$ with

$a_k(\phi) = 0$, $1 \le k \le p$ and

$$a_{p+l}(\phi) \;=\; \sum_{j=1}^{m} \mathrm{tr}\left(\Sigma_j^{-1}(\psi) X_j (X'\Sigma^{-1}(\psi)X)^{-1} X_j' \Sigma_j^{-1}(\psi) \frac{\partial \Sigma_j}{\partial \psi_l}\right) \;, \quad 1 \le l \le q \,.$$

In general, $\dot{\phi}$ may not always exist; or even if it does, may fall outside the parameter space. Of course, the MSE of $\dot{\phi}$ also may not Therefore, we consider the following *truncated* version of $\dot{\phi}$. Let Φ be the parameter space for ϕ. Let ϕ^* be a fixed vector in Φ. (In practice, ϕ^* may be a reasonable guess of the true ϕ). Let $\dot{\phi}$ be the solution to (1) if such a solution exists and lies in Φ; otherwise, let $\dot{\phi} = \phi^*$. Define the estimator $\hat{\phi}$ as follows: $\hat{\phi} = \dot{\phi}$ if $|\dot{\phi}| \le K(\log m)^\alpha$; and $\hat{\phi} = \phi^*$ otherwise, where K and α are positive (known) constants. It is clear that such a truncation will not affect the asymptotic properties such as consistency and efficiency of the estimator.

3. Jackknifing MSE of EBP. The main interest of this section is the estimation of the MSE of the proposed EBP, $\mathrm{MSE}(\hat{\theta}) = E(|\hat{\theta} - \theta|^2)$. We propose to do so by the *Jackknife* method. For such a purpose, we define the M-estimator $\dot{\phi}_{-i}$ after deleting the ith observation, i.e., $\hat{\phi}_{-i}$ is obtained likewise from the solution $\dot{\phi}_{-i}$ to the equation:

$$F_{-i}(\phi; Y_{-i}) \;=\; \sum_{j \ne i} f_j(\phi; Y_j) + a_{-i}(\phi) \;=\; 0 \,, \tag{3}$$

where $Y_{-i} = (Y_j)_{j \ne i}$. Note that $a_{-i}(\cdot)$ may not be the same function as $a(\cdot)$. Observe that

$$\mathrm{MSE}(\hat{\theta}) \;=\; E(|\hat{\theta} - \check{\theta}|^2) + E(|\check{\theta} - \theta|^2) \;=\; \mathrm{MSAE}(\hat{\theta}) + \mathrm{MSE}(\check{\theta}) \,, \tag{4}$$

where MSAE stands for "mean squared approximation error" (to the best predictor). A Jackknife estimator of the first term on the right side of (2) is given by

$$\widehat{\mathrm{MSAE}}(\hat{\theta}) \;=\; \frac{m-1}{m} \sum_{i=1}^{m} |\hat{\theta}_{-i} - \hat{\theta}|^2 \,, \tag{5}$$

where $\hat{\theta}_{-i} = \pi(Y_S; \hat{\phi}_{-i})$. Note that we keep Y_S the same (i.e., not affected by deleting the ith observation) in all $\hat{\theta}_{-i}$s. As for the second term, it is often possible to obtain a closed form expression which is a function of ϕ. Suppose $\mathrm{MSE}(\check{\theta}) = b(\phi)$. Then, a Jackknife estimator of $b(\phi)$ is given by

$$\widehat{\mathrm{MSE}}(\check{\theta}) \;=\; b(\hat{\phi}) - \frac{m-1}{m} \sum_{i=1}^{m} [b(\hat{\phi}_{-i}) - b(\hat{\phi})] \,. \tag{6}$$

Therefore, the Jackknife estimator of the MSE of $\hat{\theta}$ is

$$\widehat{\mathrm{MSE}}(\hat{\theta}) \;=\; \widehat{\mathrm{MSAE}}(\hat{\theta}) + \widehat{\mathrm{MSE}}(\check{\theta}) \,. \tag{7}$$

It can be shown that under certain regularity conditions, the bias of $\widehat{\mathrm{MSE}}(\hat{\theta})$ is of the order $o(m^{-1})$. As a by product, we also obtain the asymptotic unbiasedness of $MSE(\phi)$.

4. Mixed logistic models. Suppose that, conditional on p_{ij}, Y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are independent *Bernoulli* random variable with $P(Y_{ij} = 1 | p_{ij}) = p_{ij}$. Furthermore, suppose that conditional on the random effects $\alpha_1, \ldots, \alpha_m$, $\mathrm{logit}(p_{ij}) = x_{ij}^t \beta + \alpha_i$, where $x_{ij} = (x_{ijk})_{1 \le k \le p}$ is a vector of known covariates, β is a vector of unknown regression coefficients, and $\mathrm{logit}(t) =$

$\log(t/(1-t))$. Assume the α's are independent and distributed as $N(0,\sigma^2)$. Then, (5.1) is a special case of the generalized linear mixed models which have received considerable attention (e.g., Breslow and Clayton (1993), Lee and Nelder (1996)). Such models as (5.1) have been used in small-area inference with binary variables (e.g., Malec *et al* (1997)).

Suppose that one is interested in predicting a (possibly nonlinear) mixed effect $\theta = h_i(\beta, \alpha_i)$. For example, $\theta = \alpha_i$; or, if the covariates take values from a finite set $\{x_1, \ldots, x_K\}$. $\theta = \sum_{k=1}^K w_k \text{logit}^{-1}(x_k^t\beta + \alpha_i)$, where w_k, $1 \le k \le K$ is a set of weights, and $\text{logit}^{-1}(u) = e^u/(1+e^u)$.

Jiang and Lahiri (1998) derive the best predictor of θ as

$$\check{\theta} = E(\theta|Y) = \frac{Eh_i(\beta, \sigma\xi)\exp(\psi_i(Y_{i\cdot}, \sigma\xi, \beta))}{E\exp(\psi_i(Y_{i\cdot}, \sigma\xi, \beta))} = \pi_i(Y_{i\cdot}, \phi)\,, \tag{8}$$

where $\psi_i(k,u,v) = ku - \sum_{j=1}^{n_i}\log(1+\exp(x_{ij}^t v + u))$, $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$, $\phi = (\beta^t\ \sigma)^t$, and the expectations are taken over $\xi \sim N(0,1)$. It is also shown that $\text{MSE}(\check{\theta}) = Eh_i^2(\beta, \sigma\xi) - \sum_{k=0}^{n_i}\pi_i^2(k,\phi)p_i(k,\phi) \equiv b_i(\phi)$, where $p_i(k,\phi) = P(Y_{i\cdot} = k)$. The empirical best predictor is given by $\hat{\theta} = \pi_i(Y_{i\cdot}, \hat{\phi})$.

As for the M-estimators, we consider the method of moments (MM) estimators of Jiang (1998). The MM estimator for ϕ is the solution to the following system of equations:

$$\sum_{i=1}^m\sum_{j=1}^{n_i} x_{ijk}Y_{ij} = \sum_{i=1}^m\sum_{j=1}^{n_i} x_{ijk}E_\phi Y_{ij}\,, \quad 1 \le k \le p\,, \tag{9}$$

$$\sum_{i=1}^m\sum_{j\ne l} Y_{ij}Y_{il} = \sum_{i=1}^m\sum_{j\ne l} E_\phi Y_{ij}Y_{il}\,. \tag{10}$$

Note that $E_\phi Y_{ij} = E\text{logit}^{-1}(x_{ij}^t\beta + \sigma\xi)$, $EY_{ij}Y_{il} = E\text{logit}^{-1}(x_{ij}^t\beta + \sigma\xi)\text{logit}^{-1}(x_{il}^t\beta + \sigma\xi)$, $j \ne l$. Jiang and Lahiri (1998) showed that, under suitable conditions the MM estimators are consistent uniformly at rate m^{-d} for any $d > 0$. It follows can be shown that $E\widehat{\text{MSE}}(\hat{\theta}) = \text{MSE}(\hat{\theta}) + o(m^{-1-\epsilon})$ for any $0 < \epsilon < 1/2$.

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9-25.

Huber, P. J. (1981), *Robust Statistics*, Wiley, New York.

Jiang, J. (1996), REML estimation: asymptotic behavior and related topics, *Ann. Statist.* 24, 255-286.

Jiang, J., and Lahiri, P. (1998), Empirical best prediction for small area inference with binary data, submitted.

Lee, Y., and Nelder, J. A. (1996), Hierarchical generalized linear models (with discussion), *J. Roy. Statist. Soc.* B 58, 619-678.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), Small area inference for binary variables in the national health interview survey, *J. Amer. Statist. Assoc.* 92, 815-826.

Prasad, N.G.N., and Rao, J.N.K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163-171.

Reeds, J. A. (1978), Jackknifing maximum likelihood estimates, *Ann. Statist.* 6, 727-739.

Shao, J., and Tu, D. (1995), *The Jackknife and The Bootstrap*, Springer, New York.

Accounting for Uncertainty About Variances in Small Area Estimation

William R. Bell, Room 3000-4, U.S. Bureau of the Census, Washington, DC 20233-9100
wbell@census.gov

1. Introduction

The model of Fay and Herriot (1979) for small area estimation can be written

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m \tag{1}$$

$$= (\mathbf{x}_i'\boldsymbol{\beta} + u_i) + e_i \tag{2}$$

where the y_i are direct survey estimates of true population quantities Y_i for m small areas, the e_i are sampling errors (of the y_i) independently distributed as $N(0, v_i)$, the u_i are small area random effects (model errors) distributed *i.i.d.* $N(0, \sigma_u^2)$, the $\mathbf{x}_i'$ are $1 \times r$ row vectors of regression variables for area i, and $\boldsymbol{\beta}$ is the corresponding vector of regression parameters.

From (2), letting $\boldsymbol{\Sigma} = \text{diag}(\sigma_u^2 + v_i)$, $\boldsymbol{\beta}$ can be estimated by generalized least squares (GLS): $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ with $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$, where $\mathbf{y} = (y_1, \ldots, y_m)'$, and $\mathbf{X}$ is $m \times r$ with rows $\mathbf{x}_i'$. Then the best linear unbiased predictors (BLUPs) of the Y_i can be formed and their error variances obtained from

$$\widehat{Y}_i = h_i y_i + (1 - h_i)\mathbf{x}_i'\widehat{\boldsymbol{\beta}} \tag{3}$$

$$\text{Var}(Y_i - \widehat{Y}_i) = \sigma_u^2(1 - h_i) + (1 - h_i)^2\mathbf{x}_i'\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{x}_i \tag{4}$$

where $h_i = \sigma_u^2/(\sigma_u^2 + v_i)$. From (3), the smoothed estimate $\widehat{Y}_i$ is a weighted average of the regression prediction $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ and the direct estimate y_i. The first term in (4), $\sigma_u^2(1 - h_i)$, is the inherent prediction error variance that would result if all model parameters were known. The second term in (4) accounts for additional error due to estimating $\boldsymbol{\beta}$. Given the v_i, (4) can be augmented to reflect uncertainty about σ_u^2 using asymptotic formulas (Prasad and Rao 1990, Datta and Lahiri 1997) or a Bayesian approach (Berger 1985, pp. 190-193). When only point estimates y_i and variances v_i are available, uncertainty about sampling error variances is generally ignored, though Bell and Otto (1992) address this problem for a time series application via a Bayesian model-based approach.

This paper considers different approaches to dealing with uncertainty about σ_u^2 in the context of a particular application: estimating annual poverty rates of school-aged (5-17) children for the states of the U.S. using data from the Current Population Survey (CPS). For this problem Fay and Train (1997) developed a Fay-Herriot model for each year where, for each of $m = 51$ "states" i (including the District of Columbia as a "state"), y_i is the direct CPS estimate, Y_i the true poverty rate, and $\mathbf{x}_i$ includes a constant term and three variables derived from administrative sources. (Actually, ratios differing slightly from true poverty rates were modeled.) U.S. Internal Revenue Service income tax return files supplied two variables: an analogue to state child poverty rates and also state rates of nonfiling for income taxes. Data from the U.S. Department of Agriculture were used to develop a variable reflecting state participation rates in the food stamp poverty assistance program. In addition, $\mathbf{x}_i$ includes the residual from regressing 5-17 state poverty rates from the previous (1990) decennial census on the other regression variables for 1989 (the census income reference year). The v_i were obtained from a sampling error model of Otto and Bell (1995) that involved fitting a generalized variance function (GVF) to five years of direct variance and covariance estimates for each state produced by Fay and Train (1995). This application is an important component of the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. For information, see the SAIPE web site at http://www.census.gov/hhes/www/saipe.html.

Section 2 of this paper examines, in the context of the Fay and Train (1997) model, different approaches to dealing with uncertainty about σ_u^2 (given the v_i) and their effects on prediction error variances. Future work will explore a Gibbs sampling scheme to also recognize uncertainty about sampling error variances using the model of Otto and Bell (1995).

2. Accounting for Uncertainty About the Model Error Variance (σ_u^2)

Three estimation approaches are considered here: maximum likelihood (ML), restricted ML (REML), and the less-familiar mean likelihood (MEL). A Bayesian analysis is also explored. First, note REML maximizes the restricted likelihood (Harville 1977, p. 325)

$$L(\sigma_u^2) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \left|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})} \tag{5}$$

where $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\sigma_u^2)$ is given by GLS. Omitting the term $\left|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right|^{-1/2}$ from (5) gives the concentrated likelihood (for σ_u^2) maximized by ML. Also, (5) normalized to integrate to 1 is the Bayesian posterior density of σ_u^2 under the flat prior $p(\boldsymbol{\beta}, \sigma_u^2) \propto$ constant (Berger 1985, p. 192). The corresponding posterior mean of σ_u^2 is the same as the mean likelihood estimate (Barnard 1949).

For the Fay and Train (1997) model of U.S. 5-17 year-old state poverty rates, Table 1 shows the three estimates of σ_u^2 for 1989 1993. Focusing first on the left half of Table 1, note that the ML and REML estimates are both zero in the first four years.

Table1. Alternative Estimates of σ_u^2 for Five Years

	Updating v_i to Convergence			No Updating of v_i		
year	ML	REML	MEL	ML	REML	MEL
1989	0	0	1.7	4.6	4.9	6.1
1990	0	0	2.2	1.9	2.5	3.7
1991	0	0	1.6	0	0	1.6
1992	0	0	1.6	0	0	1.4
1993	.4	1.7	3.4	3.3	2.1	3.6

Having $\widehat{\sigma}_u^2 = 0$ has several unreasonable implications. First, it implies that if the Y_i were observed (if the CPS were a complete census every year), then the model would fit this data perfectly. (Note: The 1990 census data are not the Y_i for 1989 because of CPS-census measurement differences.) Second, since $\widehat{\sigma}_u^2 = 0$ implies $h_i = 0$ for all i, (3) implies that each $\widehat{Y}_i$ is just the regression prediction, $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$; the direct estimates y_i get no weight. Third, $\widehat{\sigma}_u^2 = 0$ implies that the first term on the right hand side of (4) is zero, and the prediction error variance comes entirely from the error in estimating $\boldsymbol{\beta}$. These results tend to look unreasonable, as will be seen later in Table 2.

Getting $\widehat{\sigma}_u^2 = 0$ for ML could motivate consideration of REML, which is intended to remove the downward asymptotic bias of ML (Datta and Lahiri 1997, p. 8). Table 1 shows that in this application, however, REML is of little help. The mean likelihood estimates, or Bayesian posterior means of σ_u^2, look more reasonable. (These were computed for Table 1 as $\int \sigma_u^2 L(\sigma_u^2) d\sigma_u^2 / \int L(\sigma_u^2) d\sigma_u^2$, with the integrations done numerically by Simpson's rule over 100 equal subintervals of $\sigma_u^2 \in [0, 20]$, an interval judged from graphs to contain essentially all the posterior probability for σ_u^2 for all years.) The reason for the differences between the estimators of σ_u^2 is easy to see from graphs (not shown) of the marginal posterior density $(L(\sigma_u^2)/\int L(\sigma_u^2) d\sigma_u^2)$, which reveal a long right tail in all years. Since the marginal posteriors are not concentrated near $\sigma_u^2 = 0$, the posterior means substantially exceed the posterior modes (mean likelihood estimators exceed REML estimators).

The estimation scheme used by Fay and Train (1997) involved iteratively updating the v_i given each new estimate of $(\boldsymbol{\beta}, \sigma_u^2)$. If superscript (k) denotes the kth iteration, the update of v_i used was $v_i^{(k)} = v_i^{(0)}[\mathbf{x}_i'\widehat{\boldsymbol{\beta}}^{(k)}(1 - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}^{(k)})]/[y_i(1 - y_i)]$, where $v_i^{(0)}$ are the original estimated v_i from the sampling error model of Otto and Bell (1995). The idea was to adjust the v_i at each iteration to be consistent with the current estimate $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. To find $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\sigma}_u^{2(k)}$ the $v_i^{(k-1)}$ were used. Convergence was effectively achieved in two iterations. For comparison, the right half of Table 1 shows the estimates of σ_u^2 without updating the v_i. The results are different in some cases, though some zero estimates for σ_u^2 still occur. For the remainder of this paper, results from updating v_i to convergence are used.

Tables 2 and 3 show some alternative prediction error variances for 1992 (when $\widehat{\sigma}_u^2 = 0$ for ML and REML) and for 1993, respectively. Also shown are CPS sample sizes n_i (number of households in the March CPS sample), CPS direct poverty rate estimates y_i, and direct sampling variance

estimates $v_i^{(0)}$. Results are shown for four states in increasing order of $v_i^{(0)}$: California (CA), the largest state with the largest sample size and lowest direct variance; North Carolina (NC); Indiana (IN); and Mississippi (MS). The tables show variances (ML^1, $REML^1$, and MEL^1) obtained by plugging $\widehat{\sigma}_u^2$ (and corresponding fully updated v_i) into (4) for $\widehat{\sigma}_u^2$ given by ML, REML, and MEL. For ML and REML the tables also show prediction error variances (ML^2 and $REML^2$) augmented as in Datta and Lahiri (1997) to asymptotically account for error in estimating σ_u^2. The tables also show two "Bayesian posterior variances" described later.

Table 2. Alternative Prediction Error Variances for Four States for 1992

state	n_i	y_i	$v_i^{(0)}$	ML^1	ML^2	$REML^1$	$REML^2$	MEL^1	$Bayes^1$	$Bayes^2$
CA	4,927	20.9	1.9	1.3	3.6	1.3	2.8	1.5	1.4	1.4
NC	2,400	23.0	5.5	.6	2.0	.6	1.2	1.6	1.4	2.0
IN	670	11.8	9.3	.3	1.4	.3	.6	1.8	1.6	1.7
MS	796	29.6	12.4	2.8	3.8	2.8	3.0	4.1	3.9	4.0

Table 3. Alternative Prediction Error Variances for Four States for 1993

state	n_i	y_i	$v_i^{(0)}$	ML^1	ML^2	$REML^1$	$REML^2$	MEL^1	$Bayes^1$	$Bayes^2$
CA	4,639	23.8	2.3	1.5	3.2	1.6	2.2	1.7	1.7	1.7
NC	2,278	17.0	4.5	1.0	2.4	1.7	2.2	2.2	2.0	2.0
IN	650	10.3	8.5	.8	1.9	1.8	2.2	2.9	2.7	3.0
MS	747	30.5	13.6	3.2	4.3	4.2	4.5	5.2	5.0	5.1

First consider ML^1 and $REML^1$ in 1992, which are the same since $\widehat{\sigma}_u^2 = 0$ for both. When $\sigma_u^2 = 0$, (4) reduces to $\text{Var}(Y_i - \widehat{Y}_i) = \mathbf{x}_i' \text{Var}(\widehat{\boldsymbol{\beta}}) \mathbf{x}_i$, and variation in (4) over states results solely from variations in the regression variables $\mathbf{x}_i$. Hence, the small values for NC and IN, despite their having smaller sample sizes and higher sampling variances than CA. In fact, many other states have values for (4) lower than that for CA. While these results would not be unexpected if we really believed $\sigma_u^2 = 0$, since $\sigma_u^2 = 0$ seems questionable so do these prediction error variances. Now comparing the ML^1 results from 1992 and 1993, we see substantial increases in 1993 for NC and IN. Similar large increases occur for many other states. In general, the differences between the ML^1 results in the two years seem overly large and not very plausible (suggesting problems particularly for 1992). The $REML^1$ results for 1993 show even more dramatic increases due to the larger REML estimate of $\widehat{\sigma}_u^2 = 1.7$ for 1993, and in contrast cast doubt on the 1993 ML^1 results.

Augmenting the ML and REML prediction variances as in Datta and Lahiri (1997) to reflect error in estimating σ_u^2 (ML^2 and $REML^2$ results) yields large increases, suggesting that ignoring this term can significantly underestimate prediction error variance. Note the largest contributions from estimating σ_u^2 go to the states with the lowest sampling variances. This makes some sense as the lower v_i is the more weight goes to the direct estimate y_i in (3) when $\sigma_u^2 > 0$, so uncertainty about σ_u^2 means more to states with fairly precise direct estimates. However, this also means that the unappealing pattern of many states having smaller prediction error variances than CA persists in ML^2 in both years and $REML^2$ in 1992.

Plugging the much larger mean likelihood estimates of σ_u^2 into (3) produces much larger prediction error variances than those from ML^1 and $REML^1$ for NC, IN, and MS (and for many other states not shown). It also yields a more intuitively appealing pattern with prediction error variance increasing with sampling variance.

Bayesian posterior variances can be computed as

$$\text{Var}(Y_i|\mathbf{y}) = E[\text{Var}(Y_i|\mathbf{y}, \sigma_u^2)] + \text{Var}[E(Y_i|\mathbf{y}, \sigma_u^2)] \tag{6}$$

where the outer expectation and variance on the right hand side are taken over the marginal posterior distribution of σ_u^2. These were computed by Simpson's rule in the same manner as the posterior means of σ_u^2 discussed above. $Bayes^1$ in Tables 2 and 3 denotes $E[\text{Var}(Y_i|\mathbf{y}, \sigma_u^2)]$, while $Bayes^2$ denotes $\text{Var}(Y_i|\mathbf{y})$. (The results shown are still conditional on the sampling variances v_i, set at their fully updated values from the mean likelihood estimation.) Note that $Bayes^1$ is fairly

close to MEL[1], i.e., averaging $\text{Var}(Y_i|\mathbf{y}, \sigma_u^2)$ over the posterior of σ_u^2 gives about the same result as evaluating $\text{Var}(Y_i|\mathbf{y}, \sigma_u^2)$ at the posterior mean of σ_u^2. Given this, $\text{Var}[E(Y_i|\mathbf{y}, \sigma_u^2)]$ = Bayes[2] − Bayes[1] can be thought of as accounting for uncertainty about σ_u^2, and as a Bayesian analogue to the term augmenting (4) to account for error in estimating σ_u^2. For most states $\text{Var}[E(Y_i|\mathbf{y}, \sigma_u^2)]$ is quite small, so $\text{Var}(Y_i|\mathbf{y})$ is close to $E[\text{Var}(Y_i|\mathbf{y}, \sigma_u^2)]$, and to MEL[1]. Note, however, the large difference between Bayes[1] and Bayes[2] for NC in 1992. This arises because the regression prediction for NC in 1992 (which varies little over the different estimates of σ_u^2) is $\mathbf{x}_i'\widehat{\boldsymbol{\beta}} = 17.7$, which differs substantially from the direct estimate $y_i = 23.0$. When such large differences between the direct estimate and the regression prediction occur, the conditional mean $(E(Y_i|\mathbf{y}, \sigma_u^2)$ as given by (3)) is sensitive to variation in σ_u^2. Hence, the posterior variances reflect this. A similar, though less pronounced effect occurs for IN in 1993 when $\mathbf{x}_i'\widehat{\boldsymbol{\beta}} = 15.3$ versus $y_i = 10.3$. Such occurrences are rare in this example, and when $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ and y_i are close, $\text{Var}[E(Y_i|\mathbf{y}, \sigma_u^2)]$ is close to zero. Note the difference in this result from the frequentist results, which do not depend in such a direct way on the realized data values.

Without overinterpreting the results from this particular example, they nonetheless show the potential difficulties for the frequentist approaches when the model error variance is estimated at or near zero. By averaging over the posterior of σ_u^2, the Bayesian approach avoids unreasonable results from fixing σ_u^2 at a single value near 0, and gives more intuitively plausible results.

References

[1] Barnard, George A. (1949), "Statistical Inference," *Journal of the Royal Statistical Society*, B, **11**, p. 116.

[2] Bell, William R. and Otto, Mark C. (1992), "Bayesian Assessment of Uncertainty in Seasonal Adjustment with Sampling Error Present," Research Report 92/12, Statistical Research Division, Bureau of the Census.

[3] Berger, James O., (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

[4] Datta, Gauri S. and Lahiri, Partha (1997), "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems," Technical Report 97-7, University of Georgia, Department of Statistics.

[5] Fay, R. E. and Herriott, R. A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, **74**, 269-277.

[6] Fay, Robert E. and Train, George F. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," American Statistical Association, Proceedings of the Section on Government Statistics.

[7] Fay, Robert E. and Train, George F. (1997), "Small Domain Methodology for Estimating Income and Poverty Characteristics for States in 1993," American Statistical Association, Proceedings of the Social Statistics Section, 183-188.

[8] Harville, David A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems (with discussion)," *Journal of the American Statistical Association*, **72**, 320-340.

[9] Otto, Mark C. and Bell, William R. (1995), "Sampling Error Modelling of Poverty and Income Statistics for States," American Statistical Association, Proceedings of the Section on Government Statistics, 160-165.

[10] Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of Mean Squared Errors of Small Area Estimators," *Journal of the American Statistical Association*, **78**, 47-59.