A Bayesian Approach to Small Domain Estimation

Kung-Jong Lui, Centers for Disease Control, Atlanta, GA
W. G. Cumberland, University of California, Los Angeles, CA

1. Introduction

In large scale sample surveys, samples designed to provide estimates for the entire population are often used secondarily to produce estimates of characteristics of subpopulations. Procedures depending on the distribution created by the sampling plan, such as simple expansion estimators of the subdomain means, are usually not applicable due to the small subdomain sample sizes. To require accurate estimates for all subdomains would necessitate a sample size too large to be met within the budgetary constraints of most population surveys.

Common approaches to small domain estimation, such as synthetic estimators (Gonzalez and Hoza 1978, and Levy 1971) or ratio-correlation estimators (Schmitt and Crosetti 1954) have the shortcoming that they are nearly impossible to evaluate with respect to the sampling plan and hence do not directly provide a measure of error for a given subdomain estimator.

The super-population model approach to this problem (Holt, Smith, and Tomberlin 1979, Lakke 1979, and Royall 1979) provides the measure of error and also gives a new avenue for exploration of this problem. In this paper, Bayesian estimators which are generalizations of the least-squares estimators of Holt, Smith and Tomberlin are proposed. The estimators derived do not depend on the sampling plan, but rather on an assumed model which reasonably describes the underlying population structure. Such estimators can easily incorporate auxiliary information from previous surveys with data from the current sample. Discussed are the effects on the performance of the estimators of model misspecification and of using inaccurate prior information. The use of simultaneous confidence intervals for several subdomains is discussed. The results are illustrated with an example using 26 health districts in Los Angeles County.

1-1 Generalized Least-Squares Estimators

We suppose that the finite population is divided into I mutually exclusive sub-areas labelled i=1,...,I for which we wish to produce estimates. Within each subdomain, units are further classified into J subgroups (for example, socio-economic class, age, etc.); these are labelled j=1,...,J. The cell sizes N_{ij} resulting from this cross-classification are assumed to be known. Let y_{ijk} be the measurement on the k^{th} individual in the ij^{th} cell and

$$T_i = \sum_{j=1}^{J} \sum_{k=1}^{N_{ij}} y_{ijk} ,$$

the total for the i^{th} subdomain. The primary focus is to estimate the T_i's.

Letting s_{ij} denote the n_{ij} sampled units in the ij^{th} cell, we use $\sum_{k\varepsilon s_{ij}} y_{ijk}$ to denote the sample sum, and $\bar{y}_{ij.}$ the average for the sample units in cell ij. Standard dot notation will be used for sample averages and sums.

Holt, Smith and Tomberlin (1979) incorporating the implicit assumptions of the synthetic estimator, $\hat{T}_i^S = \sum_j N_{ij}\bar{y}_{.j.}$, derived the modified synthetic estimator. The estimator follows from this model for the population structure:

$$y_{ijk} = \beta_j + \varepsilon_{ijk}$$

where ε's are uncorrelated with mean 0 and variance σ^2. Its form is:

$$\hat{T}_i^{MS} = \sum_j \sum_{k\varepsilon s_{ij}} y_{ijk} + \sum_j \sum_{k\notin s_{ij}} \bar{y}_{.j.}$$

with prediction variance

$$\sum_j (N_{ij}-n_{ij})\{1+(N_{ij}-n_{ij})/n_{.j}\}\,\sigma^2 .$$

The estimator $\hat{T}_i^{MS}$ cannot take advantage of the information about T_i from earlier surveys or censuses; hence it is natural to extend the model by assuming β_j to be a random variable possessing some known distribution, incorporating knowledge from previous surveys into this distribution. In the terminology of "borrowing strength" we borrow not only from the other cells of the current survey, but also from previous surveys and censuses.

We assume

$$(1.1)\quad y_{ijk} = \beta_j^* + \varepsilon_{ijk} \quad \text{where } \varepsilon_{ijk} \text{ iid } N(0,\sigma^2),\ \beta_j^* \text{ iid } N(\beta_j, \sigma_B^2)$$

and that β_j^* and ε_{ijk} are independent. Then the UMVU estimator of T_i is

$$(1.2)\quad \hat{T}_i^{GSJ} = \sum_j \sum_{k\varepsilon s_{ij}} y_{ijk} + \sum_j \sum_{k\notin s_{ij}} [(1-\lambda_j)\beta_j + \lambda_j \bar{y}_{.j.}]$$

with prediction variance

$$(1.3)\quad V(\hat{T}_i^{GSJ}-T_i) = \sum_j (N_{ij}-n_{ij})\{1+(N_{ij}-n_{ij})\lambda_j/n_{.j}\}\sigma^2$$

where $\lambda_j = n_{.j}\sigma_B^2(n_{.j}\sigma_B^2+\sigma^2)^{-1}$

$$= n_{.j}K(n_{.j}K+1)^{-1}$$

and $K = \sigma_B^2/\sigma^2$.

The estimator $\hat{T}_i^{GS1}$ results when we alter model (1.1) by making $\beta_1 = \beta_2 = \ldots = \beta_J = \beta_o$. Its prediction variance is the same as that of $\hat{T}_i^{GSJ}$. These results follow immediately from standard Bayesian techniques. (See for example, Scott and Smith, 1969). They can easily be seen to be a result of the following Lemma which is presented for future reference.

Lemma

Partition the population $\underset{\sim}{Y}' = (\underset{\sim}{Y}_s', \underset{\sim}{Y}_r')$ where $\underset{\sim}{Y}_s$ and $\underset{\sim}{Y}_r$ are the vectors of measurements for the sampled and non-sampled units respectively. Assume

$$\begin{pmatrix}\underset{\sim}{Y}_s\\ \underset{\sim}{Y}_r\end{pmatrix} = \begin{pmatrix}X_s\\ X_r\end{pmatrix}\underset{\sim}{\beta}^* + \begin{pmatrix}\underset{\sim}{\varepsilon}_s\\ \underset{\sim}{\varepsilon}_r\end{pmatrix}$$

and $\underset{\sim}{\beta}^* = A\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}^{(2)}$

where $\underset{\sim}{\beta}$ is constant, (X_s', X_r') and A are known matrices of auxiliary variables, and

$$\begin{pmatrix} \underset{\sim}{\varepsilon}_s \\ \underset{\sim}{\varepsilon}_r \end{pmatrix} \sim N\left(\underset{\sim}{0}, \begin{bmatrix} \Sigma_{ss} & \Sigma_{sr} \\ \Sigma_{rs} & \Sigma_{rr} \end{bmatrix} \right)$$

$\underset{\sim}{\varepsilon}^{(2)} \sim N(\underset{\sim}{0}, V)$ independently of $\underset{\sim}{\varepsilon}_s$ and $\underset{\sim}{\varepsilon}_r$.

Then the best predictor of

$$\underset{\sim}{\lambda}'\underset{\sim}{Y} = \underset{\sim}{\lambda}_s' \underset{\sim}{Y}_s + \underset{\sim}{\lambda}_r' \underset{\sim}{Y}_r$$

is given by

$$\underset{\sim}{\lambda}_s' \underset{\sim}{Y}_s + \underset{\sim}{\lambda}_r' E(\underset{\sim}{Y}_r | \underset{\sim}{Y}_s)$$

where $\underset{\sim}{\lambda}' = (\underset{\sim}{\lambda}_s', \underset{\sim}{\lambda}_r')$ is a vector of 0's and 1's and

$$E(\underset{\sim}{Y}_r | \underset{\sim}{Y}_s) = X_r E(\beta^* | \underset{\sim}{Y}_s) + \Sigma_{rs} \Sigma_{ss}^{-1} (\underset{\sim}{Y}_s - X_s E(\beta^* | \underset{\sim}{Y}_s))$$

$$= X_r A\beta + (X_r V X_s' + \Sigma_{rs})(X_s V X_s' + \Sigma_{ss})^{-1} (\underset{\sim}{Y}_s - X_s A\beta)$$

Proof: The expression for $E(\underset{\sim}{Y}_r | \underset{\sim}{Y}_s)$ follows from

$$E(\underset{\sim}{Y}_r | \underset{\sim}{Y}_s, \beta^*) = X_r \beta^* + \Sigma_{rs} \Sigma_{ss}^{-1} (\underset{\sim}{Y}_s - X_s \beta^*).$$

The result giving $E(\beta^* | \underset{\sim}{Y}_s)$ is an immediate consequence of writing the joint distribution of $\underset{\sim}{Y}_s$ and β^* :

$$\begin{pmatrix} \underset{\sim}{Y}_s \\ \beta^* \end{pmatrix} \sim N\left(\begin{pmatrix} X_s A\underset{\sim}{\beta} \\ A\underset{\sim}{\beta} \end{pmatrix}, \begin{bmatrix} X_s V X_s' + \Sigma_{ss} & X_s V \\ V X_s' & V \end{bmatrix} \right).$$

Note that in both $\hat{T}^{GSJ}$ and $\hat{T}^{GS1}$, a weighted average of prior information and current data is used to predict the unobserved Y's:

$$(1-\lambda_j)\beta_j + \lambda_j \bar{y}_{.j.} \quad \text{for } \hat{T}^{GSJ}$$

and $$(1-\lambda_j)\beta_o + \lambda_j \bar{y}_{.j.} \quad \text{for } \hat{T}^{GS1}.$$

These estimators become the modified synthetic estimator $\hat{T}^{MS}$ when $\kappa = \sigma_\beta^2 / \sigma^2 \to \infty$. Since λ_j is an increasing function of κ, the less accurate the prior information is (relative to the current), the larger the prediction variance becomes (see (1.3)).

Often one can provide a guess as to the value of κ, say κ_F. It is easy to see that even if κ_F is incorrect the true prediction variance of $\hat{T}^{GSJ}$ or $\hat{T}^{GS1}$ (using κ_F) is smaller than that of $\hat{T}^{MS}$ as long as $\kappa_F > \kappa/2$. Thus it is not necessary to have exact prior information regarding the variance of β^* to improve on the simple least-squares predictor $\hat{T}^{MS}$. Since an incorrect guessed value of κ_F does not bias the generalized synthetic estimators, the condition $\kappa_F > \kappa/2$ is sufficient to guarantee that $\hat{T}^{GSJ}$ and $\hat{T}^{GS1}$ are superior to $\hat{T}^{MS}$ with respect to the MSE.

If the parameters β_j (or β_o) are unknown we can substitute in the formulas for $\hat{T}^{GSJ}$ and $\hat{T}^{GS1}$ the UMVU estimators

$$\hat{\beta}_j = \bar{y}_{.j.}$$

$$\text{and} \quad \hat{\beta}_o = \Sigma\lambda_j \bar{y}_{.j.} / \Sigma\lambda_j$$

giving predictors $\hat{\hat{T}}^{GSJ}$ and $\hat{\hat{T}}^{GS1}$ respectively. Now $\hat{\hat{T}}^{GSJ}$ is identical to the least-squares estimator $\hat{T}^{MS}$ and even knowing κ makes no difference in predicting T_i. The empirical Bayes predictor $\hat{\hat{T}}^{GS1}$ is still superior to the least-squares predictor if $\kappa_F > \kappa$.

These properties of the generalized synthetic estimators are valid even when some cells have $n_{ij} = 0$. To calculate $\hat{\hat{T}}^{GSJ}$ requires that the stratum sample sizes

$$n_{.j} = \sum_i n_{ij} > 0$$

so that $\hat{\beta}_j$ can be calculated. The variances in our models need not be constant; if we allow σ^2 to vary among the different strata all that changes are the parameters λ_j.

The model used by Scott and Smith (1969) for multi-stage sampling is equivalent to (1.1) when $\beta_1 = \ldots = \beta_J = \beta_o$ are unknown, but they did not discuss the problem of subdomain estimation of T_i.

2. Generalized Expansion and Direct Estimators

For notational convenience we keep the classification into IJ cells, although for defining expansion estimators it is only necessary to have the I domains defined.

If we suppose the I domains are strata then a common estimator for T_i is the stratum expansion estimator, $\hat{T}_i^E = N_{i.} \bar{y}_{i..}$. When I is large and we have post-stratified subdomains some may have a small or zero sample size. In this situation the use of prior surveys in forming Bayes predictors is an intuitively appealing way of estimating the T_i.

2-1 Generalized Expansion Estimator

We consider the following models:

(2.1) $y_{ijk} = \alpha_i^* + \varepsilon_{ijk}$, ε_{ijk} iid $N(0, \sigma^2)$
independent of α_i^* iid $N(\alpha_i, \sigma_\alpha^2)$

(2.2) $y_{ijk} = \alpha_i^* + \varepsilon_{ijk}$, ε_{ijk} iid $N(0, \sigma^2)$
independent of α_i^* iid $N(\alpha_o, \sigma_\alpha^2)$

Application of the Lemma gives predictors which are generalizations of the expansion estimators:

$$(2.3)\quad \hat{T}_i^{GEI} = \sum_j \sum_{k \in s_{ij}} y_{ijk} + \sum_j \sum_{k \notin s_{ij}} [(1-\lambda_i)\alpha_i + \lambda_i \bar{y}_{i..}]$$

$$(2.4)\quad \hat{T}_i^{GE1} = \sum_j \sum_{k \in s_{ij}} y_{ijk} + \sum_j \sum_{k \notin s_{ij}} [(1-\lambda_i)\alpha_o + \lambda_i \bar{y}_{i..}]$$

where $\lambda_i = n_{i.}\sigma_\alpha^2 / (n_{i.}\sigma_\alpha^2 + \sigma^2)$.

The prediction variances are

$$V(\hat{T}_i^{GEI} - T_i) = V(\hat{T}_i^{GS1} - T_i)$$
$$= (N_{i.} - n_{i.})\sigma^2 + (N_{i.} - n_{i.})^2 \lambda_i \sigma^2 / n_{i.} .$$

If the α_i are not known, substituting $\hat{\alpha}_i = \bar{y}_{i..}$ in $\hat{T}_i^{GEI}$ leads to the simple expansion estimator

$$\hat{\hat{T}}_i^{GEI} = \hat{T}_i^E = N_{i.} \bar{y}_{i..}$$

with prediction variance $N_{i.}^2 (1 - n_{i.}/N_{i.})\sigma^2 / n_{i.}$ (as in Holt, Smith and Tomberlin 1979, model II). Substituting $\hat{\alpha}_o = \Sigma\lambda_i \bar{y}_{i..} / \Sigma\lambda_i$ into $\hat{T}_i^{GE1}$ gives

$$\hat{\hat{T}}_i^{GE1} = \sum_j \sum_{k \in s_{ij}} y_{ijk} + \sum_j \sum_{k \notin s_{ij}} [(1-\lambda_i)\hat{\alpha}_o + \lambda_i \bar{y}_{i..}]$$

with prediction variance

$$V(\hat{\hat{T}}_i^{GE1} - T_i) = (N_{i.} - n_{i.})\sigma^2$$
$$+ (N_{i.} - n_{i.})^2 (1-\lambda_i)^2 \sigma_\alpha^2 / \Sigma\lambda_i + (N_{i.} - n_{i.})^2 \lambda_i \sigma^2 / n_{i.} .$$

This can be shown to be smaller than the variance of the expansion estimator. Thus when α_i are nearly equal one may prefer the smaller MSE of $\hat{\hat{T}}^{GE1}$ to that of $\hat{T}_i^E$ even though the former has a bias under (2.1) when α_i are unknown.

2-2 Generalized Direct Estimator

The simple direct estimator,

$$\hat{T}_i^D = \sum_j N_{ij}\bar{y}_{ij.} \ ,$$

although intuitively appealing, is of little practical use since it will frequently be the case that at least one of the n_{ij} is zero. A Bayesian approach to generalizing this estimator assumes a distribution for the mean of each of the ij cells. We list several examples of models for this situation.

We assume

$$y_{ijk} = \mu^*_{ij} + \varepsilon_{ijk} \quad \text{where } \varepsilon_{ijk} \text{ iid } N(0,\sigma^2)$$

μ^*_{ij} and ε_{ijk} are independent and

(2.5) μ^*_{ij} iid $N(\mu_{oj}, \sigma^2_\mu)$, or

(2.6) μ^*_{ij} iid $N(\mu_{io}, \sigma^2_\mu)$, or

(2.7) μ^*_{ij} iid $N(\mu_o, \sigma^2_\mu)$ or

(2.8) μ^*_{ij} iid $N(\mu_{ij}, \sigma^2_\mu)$.

Applying the Lemma we find the general form for the Bayesian (or the empirical Bayes estimator), resulting from these models with known parameters (or when the parameters are unknown) is

$$\hat{T}_i^{GD} = \sum_j \sum_{k\varepsilon s_{ij}} y_{ijk} + \sum_j \sum_{k\notin s_{ij}} [(1-\lambda_{ij})Q_{ij} + \lambda_{ij}\bar{y}_{ij.}]$$

where Q_{ij} is the known prior mean $E(\mu^*_{ij})$ (or its corresponding estimate under the model).

We can see when $n_{ij}=0$ that $\lambda_{ij}=0$ so that we use Q_{ij} to predict the non-sampled individuals in cell ij and if $n_{ij} > 0$ we use a weighted average of Q_{ij} and the cell sample means.

3. Covariate Models

There are situations in which the average value of the elements in a subgroup is thought to be a function of some known auxiliary variables. Here we shall discuss several different models which are related to those discussed in the previous sections, but which incorporate the auxiliary information.

3-1 Covariate Models Related to the Synthetic Estimator

As a special case of model (1.1) we consider the situation where $E\beta^*_j$ is a linear function of a known covariate vector $\underset{\sim}{X}_j$, of dimension $p < J$. That is, we assume for $i=1,2,\ldots,I$, $j=1,2,\ldots,J$, $k=1,2,\ldots,N_{ij}$ that

(3.1) $y_{ijk} = \beta^*_j + \varepsilon_{ijk}$ where ε_{ijk} iid $N(0,\sigma^2)$

β^*_j iid $(\alpha_o + \underset{\sim}{\alpha}'\underset{\sim}{X}_j, \sigma^2_B)$,

β^*_j, ε_{ijk} are independent,

and $\underset{\sim}{X}_j = (X_{j1}, X_{j2},\ldots, X_{jp})'$ are known.

If α_o and $\underset{\sim}{\alpha}$ are known, the best estimator $\hat{T}_i^{CSP}$ of T_i under model (3.1) can be obtained by substituting $\alpha_o + \underset{\sim}{\alpha}'X_j$ for β_j in $\hat{T}_i^{GSJ}$ (see (1.2)); its prediction variance is the same as $V(\hat{T}_i^{GSJ} - T_i)$ given by (1.3). If α_o and $\underset{\sim}{\alpha}$ are unknown, one uses the estimates

$$\hat{\alpha}_o = \bar{Y}_w - \hat{\underset{\sim}{\alpha}}'\bar{\underset{\sim}{X}} \quad \text{and}$$

(3.2) $$\hat{\underset{\sim}{\alpha}} = (X_D' \wedge X_D)^{-1} X_D' \wedge \bar{\underset{\sim}{Y}}_A$$

$$\bar{Y}_w = \sum_j \lambda_j \bar{y}_{.j.} / (\sum_j \lambda_j)$$

where $X_D = \left(X_{j\iota} - \bar{X}_\iota\right)_{J\times p}$, $\bar{\underset{\sim}{X}} = \Sigma\lambda_j\underset{\sim}{X}_j/\Sigma\lambda_j$

$\Lambda = \left(\text{diag}(\lambda_j)\right)_{J\times J}$, $\bar{\underset{\sim}{Y}}_A' = \left(\bar{Y}_{.1.},\ldots,\bar{Y}_{.j.}\right)$.

Thus, the UMVU estimator, T_i^{CSP}, is given by

(3.3) $$\hat{T}_i^{CSP} = \sum_j \sum_{k\varepsilon s_{ij}} y_{ijk} + \sum_j \sum_{k\notin s_{ij}} [(1-\lambda_j)(\hat{\alpha}_o + \hat{\underset{\sim}{\alpha}}'\underset{\sim}{X}_j) + \lambda_j\bar{y}_{.j.}]$$

with the prediction variance given by

(3.4) $$V(\hat{T}_i^{CSP} - T_i) = \sigma^2 \Bigg[\sum_j (N_{ij} - n_{ij}) + \sum_j (N_{ij} - n_{ij})^2 \lambda_j / n_{.j} + \left[\sum_j (N_{ij} - n_{ij})(1-\lambda_j)\right]^2 \kappa / \sum_j \lambda_j + \left[\underset{\sim}{L}_i' \left(X_D' \wedge X_D\right)^{-1} \underset{\sim}{L}_i\right] \kappa \Bigg]$$

where

$$\underset{\sim}{L}_i' = \Big[\sum_j (N_{ij} - n_{ij})(1-\lambda_j)(X_{j1} - \bar{X}_1), \ldots, \sum_j (N_{ij} - n_{ij})(1-\lambda_j)(X_{jp} - \bar{X}_p)\Big] .$$

In particular, when p=1 we can easily see that

$$E(\hat{T}_i^{GS1} - T_i)^2 < E(\hat{T}_i^{CS1} - T_i)^2$$

if and only if $V(\hat{\alpha})/(E(\hat{\alpha}))^2 = CV^2(\hat{\alpha}) > 1$.

Thus, $\hat{T}_i^{GS1}$ though biased, can be a better (with respect to the MSE) estimator of T_i than the UMVU estimator $\hat{T}_i^{CS1}$ under model (3.1) when p=1; this occurs when the square of coefficient of variation, $CV^2(\hat{\alpha}) > 1$. This result suggests that we should be careful in deciding whether we want to include the auxiliary variable in our estimator for T_i or not.

3-2 Covariate Models Related to the Simple Direct Estimator

Following similar ideas as those in the previous section, we can incorporate the auxiliary information into the estimator for T_i under models related to the simple direct estimator. For simplicity, we restrict our discussion to only one covariate here. The estimator for the P-variate case can be derived using similar arguments.

Consider models which relate μ_{ij} to a covariate z_{ij} measured for each cell. We assume for $i=1,2,\ldots I$, $j=1,2,\ldots J$, $k=1,2,\ldots N_{ij}$ that

$$y_{ijk} = \mu^*_{ij} + \varepsilon_{ijk} \quad \text{where } \varepsilon_{ijk} \text{ iid } N(0,\sigma^2)$$

μ^*_{ij} and ε_{ijk} are independent, and

(3.5) μ^*_{ij} iid $N\left(\alpha+\beta z_{ij}, \sigma^2_\mu\right)$ or

(3.6) μ^*_{ij} iid $N\left(\beta_j z_{ij}, \sigma^2_\mu\right)$.

The UMVU estimator of T_i under model (3.5) or (3.6) can be obtained easily whether the prior parameters α, β under model (3.5) or β_j under model (3.6) are known or unknown.

We mention that the use of the model (3.6) under certain situations (for example, in which the sampled units are groups of households, but the average value per household of the characteristic measured is the same within a given j^{th} stratum) is potentially more realistic than the model (3.5). Continuing this example, suppose one wishes to estimate characteristics based on a sample of blocks, such as the number of school-aged children, or the number of dilapidated housing units, for the small areas. For these measurements, the number per block will be directly related to the number of households per block. Intuitively, we expect that the greater the number of households in a block, the greater the number of school-aged children (or the number of dilapidated housing units). The slope of β_j in the model (3.6) can be interpreted as the expected number of school-aged children per household in the j^{th} stratum and therefore $\beta_j z_{ij}$ where z_{ij} is the number of households in the ij^{th} cell, is the expected number of school-aged children in the ij^{th} cell.

Note that if z_{ij} are equal, model (3.6) will be the same as (2.5), and therefore the corresponding estimators under these two models will be identical.

4. Variance Estimation

Recall that the formulae for the prediction variance of the Bayesian estimators related to the synthetic estimator involve σ^2 and σ_B^2. Since it is often possible to guess the relative size $K = \sigma_B^2/\sigma^2$, we need only to estimate one of the parameters σ_B^2 and σ^2, and then find the other from the ratio. We thus rewrite the prediction variance for the generalized synthetic estimator as σ^2 times a function of K, N_{ij} and n_{ij}. Assuming that K is known simplifies the problem of component variance estimation and also allows us to concentrate on estimating σ^2 rather than σ^2 and σ_B^2.

Following the above idea, we assume that we know the value of K from a previous survey or one's own subjective confidence for the relative sizes of σ_B^2 and σ^2. We can easily get the UMVU estimator of σ^2 and σ_B^2 under the assumed models. For example, the UMVU estimator of σ^2 under model (1.1) when the β_j are unknown, is

$$(4.1) \qquad \hat{\sigma}^2_{GSJ} = \sum_j \sum_i \sum_{k \varepsilon s_{ij}} (y_{ijk} - \bar{y}_{.j.})^2 / (n_{..} - J)$$

which is independent of K. The UMVU estimator of σ_B^2 in this case is $K\,\hat{\sigma}^2_{GSJ}$.

4-1 Simple Interval Estimates Under the Generalized Synthetic Model with J Prior Means

Assuming model (1.1) in which β_j are known, we get the estimators $\hat{T}_i^{GSJ}$, i=1,2,...I as given in (1.2) with prediction variances given in (1.3). Let the error vector be

$$\hat{\underset{\sim}{E}}^{GSJ} = \left(\hat{T}_1^{GSJ} - T_1, \ldots, \hat{T}_I^{GSJ} - T_I\right)'.$$

Then $\hat{\underset{\sim}{E}}^{GSJ} = W \underset{\sim}{Y} + \underset{\sim}{\omega}$, a linear function of the elements of $\underset{\sim}{Y}$. The $I \times N_{..}$ matrix W is a function of n_{ij}, N_{ij}, and the λ_j, while the vector $\underset{\sim}{\omega}$ has elements $\sum_j (N_{ij} - n_{ij})(1-\lambda_j)\beta_j$.

From this it follows that $\hat{\underset{\sim}{E}}^{GSJ} \sim N(\underset{\sim}{0}, \sigma^2 C)$ where C is an $I \times I$ matrix with

$$c_{ii} = \sum_j (N_{ij} - n_{ij}) + \sum_j (N_{ij} - n_{ij})^2 \lambda_j / n_{.j} \qquad \text{and}$$

$$c_{ii'} = \sum_j (N_{ij} - n_{ij})(N_{i'j} - n_{i'j})(1-\lambda_j) K \quad \text{for } i \neq i'.$$

Note that C is of full rank, since W is of rank I and $COV(\underset{\sim}{Y})$ of rank $N_{..}$. It is easy to show that $n_{..}\hat{\sigma}^2_{GSJK}/\sigma^2$ is distributed as chi-square with $n_{..}$ degrees of freedom, where

$$\hat{\sigma}^2_{GSJK} = (\underset{\sim}{Y}_s - X_s \beta)' V_s^{-1} (\underset{\sim}{Y}_s - X_s \beta), \text{ and}$$

$\sigma^2 V_s$ is the $COV(\underset{\sim}{Y}_s)$. Further, $\hat{\underset{\sim}{E}}^{GSJ}$ and $\hat{\sigma}^2_{GSJ}$ are independent, hence from standard theory, we find

$$\hat{T}_i \pm t \left(c_{ii}\, \hat{\sigma}^2_{GSJK}\right)^{1/2},$$

where t is the $\alpha/2$ upper percentile of Student distribution with $n_{..}$ degrees of freedom, gives a $1-\alpha$ confidence interval for T_i.

4-2 Simultaneous Interval Estimates

In small domain estimation, we are frequently more interested in giving simultaneous confidence intervals for all T_i than in giving one-at-a-time confidence intervals. We present the following three methods for constructing simultaneous confidence intervals for T_i. All three methods lead to confidence intervals of the form

$$\hat{T}_i \pm h \left(c_{ii}\, \hat{\sigma}^2\right)^{1/2}, \quad i=1,2,\ldots I,$$

where h is determined by the method and α.

The simplest method, based on Bonferroni inequality is to use for h, the upper $\alpha/2I$ percentile of the t-distribution rather than $\alpha/2$ in calculating each confidence interval.

The second method is the multivariate-t method (see for example Graybill (1976)) in which to give $1-\alpha$ simultaneous confidence intervals for T_i, we use for h the $\alpha/2$ upper percentile of the standard multivariate-t distribution.

The third method, related to Scheffé's method for confidence intervals, is to use

$$h = \left(I\, F_\alpha(I, n_{..})\right)^{1/2}$$

to give the interval estimates. This can be easily derived from noting that

$$\hat{\underset{\sim}{E}}'^{GSJ}\, C^{-1}\, \hat{\underset{\sim}{E}}^{GSJ} / (I\, \hat{\sigma}^2_{GSJK}) \sim F(I, n_{..})$$

and $$\max_{\underset{\sim}{l}} \left(\underset{\sim}{l}'\, \hat{\underset{\sim}{E}}^{GSJ}\right)^2 / (\underset{\sim}{l}'\, C\, \underset{\sim}{l}) = \hat{\underset{\sim}{E}}'^{GSJ}\, C^{-1}\, \hat{\underset{\sim}{E}}^{GSJ},$$

when the max is over the I-dimension Euclidean space, excluding $\underset{\sim}{0}$. Notice that using this method not only provides simultaneous confidence intervals for all T_i, but also for all linear combinations of T_i. Therefore, this method is especially useful when we are also interested in calculating confidence intervals for $\sum_{i \varepsilon g} T_i$, where g is any collection of labels from $\{1,2,\ldots I\}$. Scheffé's method always gives longer confidence intervals for T_i than using the above multivariate-t method.

If the β_j under model (2.2) are unknown, then arguments, similar to those when the β_j are known, can show that:

$$c_{ii} = \sum_j (N_{ij} - n_{ij}) + \sum_j (N_{ij} - n_{ij})^2 / n_{.j}$$

and

$$c_{ii'} = \sum_j (N_{ij} - n_{ij})(N_{i'j} - n_{i'j}) / n_{.j}.$$

We then get the simultaneous confidence intervals as $\hat{T}_i \pm h \left(c_{ii} \hat{\sigma}^2_{GSJ} \right)^{1/2}$ where h is chosen with degrees of freedom $n_{..} - J$ from the appropriate table corresponding to the multiple-t, multivariate-t or Scheffé's techniques.

5. An Example Using Los Angeles County Health Districts

The results of the preceding sections show that the Bayesian approach to small domain estimation has the potential to be an valuable tool in the hands of the practitioner. The ability to "borrow strength" not only from the current survey, but also from previous studies and censuses has great appeal. The ease with which one can get variance estimators and with which one can take advantage of existing procedures for simultaneous confidence interval estimation make these methods attractive. Of concern to the user of these statistics, however, is their robustness to failure of the model assumptions. We showed that the Bayesian estimators do exhibit some robustness to model failure; they may in fact do rather well even when there are significant departures from the assumptions. Thus it is important to see how well these procedures perform when applied to data from actual finite populations as compared to the traditional methods in small domain estimation.

Here we present the results of an simple study in which we compared the performance of the estimators derived in section 1-2 with that of more traditional estimators. Samples of different sizes were taken from a population consisting of the 1275 census tracts of Los Angeles County, based on the 1960 census. Each census tract belonged to exactly one of 26 health districts of L.A. county; these became the small areas. The response variable, y, chosen for this study was the number of physicians in each census tract; hence T_i was the number of physicians for the i^{th} health district. This population was further classified into four strata, defined by the median house price from the 1960 census: $<$ \$10,000, \$10,001-20,000, \$20,001-30,000, and $>$ \$30,000. As a result, the population was cross-classified into 104 cells. The number of tracts per cell, N_{ij}, ranged from 0 to 56.

With the tracts themselves as sampling units, we used a simple random sampling plan to select our samples. The traditional estimators used in this study were the synthetic estimator, $\hat{T}^S_i$, and the expansion estimator, $\hat{T}^E_i$. The direct estimator was not used since it is undefined when a cell has zero sample units. Three estimators, related to the synthetic estimator, were chosen for comparison: $\hat{T}^{GS1}_i$, $\hat{T}^{GSJ}_i$, and $\hat{T}^{GSJ}_i$. These estimators are sensible choices for this population since is it reasonable to relate the number of physicians to the tract median house value. Assuming that within a stratum, the number of physicians per census tract is relatively constant leads one to consider model (1.1) as representative of the structure of this population.

Since the estimators $\hat{T}^{GS1}_i$ and $\hat{T}^{GSJ}_i$ depend on the value of K, several different values of K were chosen. In this study we used 1,2,3,4,5, and 6, which cover most values of interest. From a preliminary study, a value of K greater than 6 led to these three estimators being almost identical. Values of K less than 1 were not considered of practical interest. Since the prior mean needed to calculate $\hat{T}^{GSJ}_i$ is not known for this population we chose to use the population column mean for β_j -- this represents having the best possible information on them, and hence we could evaluate how much efficiency is lost when one must estimate these parameters from the sample data.

To measure the performances of the estimators, we used criteria that have been commonly used to evaluate traditional small domain estimators (see for example Levy (1971), Schaible (1979), O'Hare (1976), Gonzalez (1979), and Crosetti and Schmitt (1956)). These include the root mean squared error (RMSE),

$$\left[\sum_{i=1}^{I} (\hat{T}_i - T_i)^2 / I \right]^{1/2},$$

the sample correlation coefficient (CORR), and the mean of percentage absolute difference (MPAD),

$$\sum_{i=1}^{I} |\hat{T}_i - T_i| / (I\hat{T}_i).$$

Four different sample sizes, 255, 128, 64, and 39 were used. One sample was taken for each sample size and each value of K. Note that when $n_{i.} = 0$, $\hat{T}^E_i$ is not defined, and therefore, RMSE and CORR for $\hat{T}^E_i$ were calculated only based on the subareas in which there was at least one sampled unit. Furthermore, the MPAD is undefined, if for any subarea i, $\hat{T}^E_i$ is zero. The results are given in Table 1.

Note that the simple expansion estimator was the poorest among these five estimators, especially when the sample sizes were small. For n=255, the three Bayesian estimators were not very different and did not seem to depend on the ratio K. These results are concordant with the fact that when n is large, λ_j is close to 1 and hence these estimators depend little on the prior knowledge. When the sample size was moderate, for example 128, the estimator $\hat{T}^{GSJ}_i$ was as good as the estimator $\hat{T}^{GSJ}_i$ regardless of K. However, when the sample size decreased to 64 or even less to 39, the estimator $\hat{T}^{GSJ}_i$ was the best estimator among these five with respect to the root-mean squared error. This is expected since using very accurate information in the Bayesian estimator is especially useful when the sample size is small. In our example, $\hat{T}^{GSJ}_i$ and $\hat{T}^{GSJ}_i$ consistently performed better than either the synthetic or the simple expansion estimator. A comparison of $\hat{T}^{GS1}_i$ with $\hat{T}^S_i$, showed that $\hat{T}^{GS1}_i$ was better when the sample size was large, (255 or 128 in our example), while $\hat{T}^S_i$ performed as well as $\hat{T}^{GS1}_i$ when the sample size was small (39 in our example) and K = 1 or 2. Note that the above cases, for K > 2, $\hat{T}^{GS1}_i$ was still preferable to $\hat{T}^S_i$ with respect to the root-mean squared error. This suggests that if we are not sure of the true value of K, we are better off using too large rather than too small a value of K, especially when the sample size is small.

6. Conclusions

Our example, though limited in scope, shows that using a Bayesian approach to derive small domain estimators can often give a dramatic improvement over traditional estimators. Bayesian estimators are generally more flexible in borrowing information from related areas that the synthetic estimator. The wide choice of models allows a practitioner to take advantage of

whatever information he has at hand about the population; he can, for example, easily incorporate auxiliary information into the estimators. Further, one can get a measure of error, which allows the construction of simultaneous confidence intervals for T_i for each small area. The theoretical and empirical results suggest these estimators deserve serious consideration for use in small domain estimation.

Table 5-1
Comparisons of the Performances of Different Estimators
Using L.A. County 26 Health districts

		$\hat{T}_i^{GS1}$	$\hat{T}_i^{GSJ}$	$\hat{T}_i^{GSJ}$	$\hat{T}_i^{S}$	$\hat{T}_i^{E}$
		(Sample Size = 255)				
	RMSE	105.45	105.53	105.51	147.59	250.18
K=1	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.30	0.29	0.25	0.32	0.67
	RMSE	105.40	105.52	105.51	147.59	250.18
K=2	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.29	0.29	0.25	0.32	0.67
	RMSE	105.42	105.52	105.51	147.59	250.18
K=3	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.29	0.29	0.25	0.32	0.67
	RMSE	105.43	105.52	105.51	147.59	250.18
K=4	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.29	0.29	0.25	0.32	0.67
	RMSE	105.45	105.51	105.51	147.59	250.18
K=5	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.29	0.29	0.25	0.32	0.67
	RMSE	105.45	105.51	105.51	147.59	250.18
K=6	CORR	0.97	0.97	0.97	0.94	0.95
	MPAD	0.29	0.29	0.25	0.32	0.67
		(Sample Size = 128)				
	RMSE	137.11	130.48	131.95	155.31	562.85
K=1	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.34	0.33	0.29	0.34	*
	RMSE	134.51	131.16	131.95	155.31	562.85
K=2	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.34	0.33	0.29	0.34	*
	RMSE	133.65	131.41	131.95	155.31	562.85
K=3	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.34	0.33	0.29	0.34	*
	RMSE	133.22	131.54	131.95	155.31	562.85
K=4	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.34	0.33	0.29	0.34	*
	RMSE	132.96	131.64	131.95	155.31	562.85
K=5	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.33	0.33	0.29	0.34	*
	RMSE	132.79	131.67	131.95	155.31	562.85
K=6	CORR	0.96	0.96	0.96	0.94	0.92
	MPAD	0.33	0.33	0.29	0.34	*
		(Sample Size = 64)				
		$\hat{T}_i^{GS1}$	$\hat{T}_i^{GSJ}$	$\hat{T}_i^{GSJ}$	$\hat{T}_i^{S}$	$\hat{T}_i^{E}$
	RMSE	229.98	199.11	218.25	255.11	636.59
K=1	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.40	0.40	0.42	*
	RMSE	224.55	207.91	218.25	225.11	636.59
K=2	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.42	0.40	0.42	*
	RMSE	222.56	211.17	218.25	225.11	636.59
K=3	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.42	0.40	0.42	*
	RMSE	221.52	212.87	218.25	225.11	636.59
K=4	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.43	0.40	0.42	*
	RMSE	220.89	213.91	218.25	225.11	636.59
K=5	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.43	0.40	0.42	*
	RMSE	220.46	214.61	218.25	225.11	636.59
K=6	CORR	0.96	0.96	0.96	0.95	0.87
	MPAD	0.44	0.43	0.40	0.42	*
		(Sample Size = 39)				
	RMSE	221.11	172.93	212.42	216.40	705.61
K=1	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.42	0.38	0.39	0.41	*
	RMSE	216.67	188.67	212.42	216.40	705.61
K=2	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.43	0.39	0.39	0.41	*
	RMSE	215.31	195.50	212.42	216.40	705.61
K=3	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.43	0.40	0.39	0.41	*
	RMSE	214.58	199.29	212.42	216.40	705.61
K=4	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.43	0.41	0.39	0.41	*
	RMSE	214.14	201.70	212.42	216.40	705.61
K=5	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.43	0.41	0.39	0.41	*
	RMSE	213.85	203.36	212.42	216.40	705.61
K=6	CORR	0.95	0.96	0.95	0.95	0.88
	MPAD	0.43	0.42	0.39	0.41	*

* The value is undefined.

REFERENCES

Crosetti, A.H. and Schmitt, R.C. (1956), "A Method of Estimating the Intercensal Population of Counties," Journal of the American Statistical Association, 51, 587-590.

Gonzalez, M.E. (1979) "Case Studies on the Use and Accuracy of Synthetic Estimates: Unemployment and Housing Applications," in Synthetic Estimates for Small Areas Statistical Workshop Papers and Discussion, Ed. J. Steinberg, NIDA Research Monograph 24, 142-153.

Gonzales, M.E., and Hoza, C. (1978), "Small Area Estimation with Applications to Unemployment and Housing Estimates," Journal of the American Statistical Association, 73, 7-15.

Graybill, Franklin A. (1976), Theory and Application of the Linear Model, Duxbury Press, North Scituate Massachusetts.

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

Lakke, P. (1979), "A Predictive Approach to Subdomain Estimation in Finite Populations," Journal of the American Statistical Association, 74, 355-358.

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

O'Hare, W. (1976), "Report on a Multiple Regression Method for Making Population Estimates," Demography, 13, 369-379.

Royall, R.M. (1979), "Prediction Models in Small Area Estimation," in Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion, Ed. J. Steinberg, NIDA Research Monograph 24, 63-83.

Schaible, W.L. (1979), "A Composite Estimator for Small Area Statistics," in Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion, Ed. J. Steinberg, NIDA Research Monograph 24, 142-153.

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

Schmitt, R.C., and Crosetti, A.H. (1954), "Accuracy of the Ratio-correlation Method for Estimating Postcensal Population," Land Economics, 30, 279-281.

United States Bureau of the Census. 1960 Census of Housing, Block Statistics Los Angeles County.

Los Angeles Health Districts 1960, Western Data Processing Center, Graduate School of Business Administration, University of California, Los Angeles.