

Session 11  
SMALL AREA ESTIMATION

# Small Area Estimation for the National Health Interview Survey Using Hierarchical Models

DONALD MALEC

National Center for Health Statistics; Hyattsville, MD

J. SEDRANSK

State University of New York; Albany, NY

## 1. Introduction

There is a continuing need to assess health status, practices and resources at both the national level and subnational levels. Estimates of these health items help determine the demand for quality health care and the access individuals have to it. Although NCHS survey data systems can provide much of this information at the national level, little can be provided directly at the subnational level, except for a few large states and metropolitan areas. The need for State and substate health statistics exists, however, because health and health care characteristics are known to vary geographically. Also, health care planning often takes place at the state and county level.

Using a hierarchical model, our focus is on the development of state estimators using data from the National Health Interview Survey (NHIS). Information on health status, practices and resources is collected annually in the NHIS and direct national estimates of these items are also produced annually. The NHIS is a multistage, personal interview sample survey. The current sample design uses 1,983 primary sampling units (PSU's), each PSU consisting of a single county or a group of contiguous counties (minor civil divisions are used instead of counties in New England and Hawaii). The population of 1,983 PSU's is stratified and approximately 200 are sampled with probability roughly proportional to their population sizes. Within each sampled PSU clusters of households are formed and sampled. Areas within a PSU with a high concentration of blacks are oversampled. The NHIS is a cross-sectional survey: each year, a new sample containing approximately 50,000 households and 120,000 individuals is selected. (For additional details about the design of the NHIS see Massey et al. 1989.) Although the total sample size is large, the sample size in most states is too small to produce direct estimates that are sufficiently precise.

Malec and Sedransk (1985) have described Bayesian methodology appropriate for the analysis of some multi-stage sample surveys when the variables are normally distributed. We have extended this methodology to accommodate binary random variables, the predominant variables in the NHIS. Our model is similar to that of Wong and Mason (1985). However, the objective in Wong and Mason (1985) is inference about parameters in the model rather than finite population quantities. While Dempster and Tomberlin (1980) investigate small area estimation

methods for binary random variables they, like Wong and Mason, provide an empirical Bayes rather than a fully Bayes solution. Since empirical Bayes procedures often account for only a fraction of the error correctly represented in a fully Bayes approach, we prefer the latter. Recent advances in numerical methods (e.g., the Gibbs sampler) permit the employment of a full Bayesian analysis; see, e.g., Gatsonis, et al. (1993), Malec and Sedransk (1993a), and Malec, Sedransk, and Tompkins (1993).

The notation and model are described in Section 2 while the estimation methodology is presented in Section 3. Section 4 describes the techniques for fitting the proposed models, and displays the final model using data from the 1987 NHIS on utilization of physician care. There is a comparison of alternative estimators in Section 5, and evaluation of the proposed methodology is described in Section 6.

## 2. Model Specification

The model in (2.1), (2.2) and (2.3) below includes the most important features of the sample design. Our objective is to produce accurate point estimates and appropriate measures of variability by accounting for geographic variability of the response and using available covariate information.

Let  $Y_{ij}$  denote a binary random variable for individual  $j$  in county  $i$  where  $i=1, \dots, L$  and  $j=1, \dots, N_i$ . Within county  $i$  and conditional on the  $p_{ij}$ , the  $Y_{ij}$  are assumed to be independent Bernoulli random variables; i.e.,

$$Pr(Y_{ij} = y_{ij}) = p_{ij}^{y_{ij}} (1 - p_{ij})^{1 - y_{ij}}, y_{ij} \in \{0, 1\}. \quad (2.1)$$

Given the vector of  $M$  covariates corresponding to individual  $j$ ,  $\mathbf{X}_{ij}^t = (X_{ij1}, \dots, X_{ijM})$ , and  $\beta_i$ , it is assumed that

$$\ln\{p_{ij} / (1 - p_{ij})\} = \mathbf{X}_{ij}^t \beta_i. \quad (2.2)$$

To allow for the possibility of a linear regression between each element of  $\beta_i$  and a set of covariates,  $\mathbf{Z}_i^t = (Z_{i1}, \dots, Z_{ic})$ , available at the county level, assume

$$\beta_i \sim N(G_i \eta, \Gamma) \quad (2.3)$$

where, conditional on  $\eta$  and  $\Gamma$ , the  $\beta_i$ 's are independent and  $G_i = \text{Diag}(g_{i1}^t, g_{i2}^t, \dots, g_{iM}^t)$  and  $g_{ij}^t$  is a row vector of dimension  $c_j$  containing a subset of covariates from  $\mathbf{Z}_i^t$ . Additionally,  $\eta^t = (\eta_{11}, \dots, \eta_{1c_1}, \eta_{21}, \dots, \eta_{2c_2}, \dots, \eta_{M1}, \dots, \eta_{Mc_M})$ , conforming to the dimensions of  $G_i$ , and  $\Gamma$  is an  $M \times M$  positive definite matrix. Finally, reference prior distributions are assigned to  $\eta$  and  $\Gamma$ ; i.e.,

$$p(\eta, \Gamma) \propto \text{constant}. \quad (2.4)$$

Taking  $\Gamma=0$  provides a specification that is consistent with the basic assumption of synthetic estimation. In the following "synthetic estimation" refers to the use of (2.1), (2.2) with  $\underline{\beta}_i = G_i \underline{\eta}$ , and  $p(\underline{\eta}) \propto \text{constant}$ .

We include as variables in (2.2) those individual level characteristics that provide the best prediction for  $p_{ij}$ , and are reliably estimated at the county level in non-census years. Candidate variables at the county level (i.e.,  $G_i$  in (2.3)) include the variables used to define the NHIS strata (defined at the county level). By predicting for demographic groups within counties and then weighting by postcensal population estimates, estimates are automatically weighted to fixed population totals.

### 3. Estimation Methodology

#### 3.1 Bayesian Predictive Inference

In this paper, our objective is to make inference about finite population means. By first summing the  $Y_{ij}$ 's within a county and then within a state, the population mean within a state can be expressed as

$$\theta = \sum_i \sum_{j=1}^{N_i} Y_{ij} / N. \quad (3.1)$$

Formula (3.1) can represent either a mean for the entire state or for a subpopulation. The first sum is over the collection of counties within the state, while  $N_i$  is the size of the population or subpopulation in county  $i$ . Here,  $\sum_i N_i = N$ .

In (2.2) we use the variables age, sex and race because these are the only variables for which reliable estimates are available at the county level for non-Census years. In this case, (3.1) can be simplified. Suppose that in the population there are  $K$  different values of the vector  $\underline{X}_{ij}$ . Then write  $\underline{X}_{ij} = \underline{X}(k)$  for all  $ij$  having pattern  $k$  ( $k=1, \dots, K$ ).

From (3.1),

$$\theta = \sum_{(i,j) \in S} \frac{Y_{ij}}{N} + \sum_i \sum_{k=1}^K \left( \frac{N_i(k) - n_i(k)}{N} \right) \bar{Y}_{ik}^{(ns)} \quad (3.2)$$

where  $N_i(k)$  and  $n_i(k)$  are, respectively, the population and sample sizes in county  $i$  with  $\underline{X}_{ij} = \underline{X}(k)$  and  $\bar{Y}_{ik}^{(ns)}$  is the mean of the nonsampled individuals with demographic characteristic  $k$  in county  $i$ .

Letting  $\underline{y}_s$  denote the vector of sample observations, we emphasize the first two moments of  $\theta$ ,  $E(\theta | \underline{y}_s)$  and  $\text{Var}(\theta | \underline{y}_s)$ .

From (3.2),

$$E(\theta | \mathbf{y}_s) = \sum_{(i,j) \in s} \frac{y_{ij}}{N} + \sum_{i,k=1}^K \left( \frac{N_i(k) - n_i(k)}{N} \right) E(p_{ik} | \mathbf{y}_s) \quad (3.3)$$

where  $p_{ik} = \exp\{X^t(k) \beta_i\} / [1 + \exp\{X^t(k) \beta_i\}]$

and

$$\begin{aligned} \text{Var}(\theta | \mathbf{y}_s) &= \sum_{i,k=1}^K \{N_i(k) - n_i(k)\} E\{p_{ik}(1-p_{ik}) | \mathbf{y}_s\} / N^2 \\ &+ \text{Var}\left(\sum_{i,k=1}^K \{N_i(k) - n_i(k)\} p_{ik} | \mathbf{y}_s\right) / N^2. \end{aligned} \quad (3.4)$$

### 3.2 Numerical Evaluation

Since the posterior moments of  $\theta$  are nonlinear functions of  $\{\beta_i: i=1, \dots, L\}$ , and the posterior distribution,

$$f(\{\beta_i: i=1, \dots, L\}, \eta, \Gamma | \mathbf{y}_s), \quad (3.5)$$

cannot be expressed in a simple form, numerical evaluation is needed. We generate from (3.5)  $R$  sets of parameters,  $\Omega = \{\Omega^{(r)}: r=1, \dots, R\}$ , where  $\Omega^{(r)} = \{\{\beta_i^{(r)}: i=1, \dots, L\}, \eta^{(r)}, \Gamma^{(r)}\}$ . Then we evaluate the  $p_{ij}^{(r)}$  using (2.2), and obtain estimates of  $E(\theta | \mathbf{y}_s)$  and  $\text{Var}(\theta | \mathbf{y}_s)$ ,

$$\hat{E}_\theta = \sum_{(i,j) \in s} y_{ij} / N + R^{-1} \sum_{r=1}^R \left[ \sum_{(i,j) \in s} p_{ij}^{(r)} / N \right] \quad (3.6)$$

and

$$\begin{aligned} \hat{V}_\theta &= R^{-1} \sum_{r=1}^R \left[ \sum_{(i,j) \in s} p_{ij}^{(r)} (1 - p_{ij}^{(r)}) + \left\{ \sum_{(i,j) \in s} p_{ij}^{(r)} \right\}^2 \right] / N^2 \\ &- \left\{ R^{-1} \sum_{r=1}^R \left[ \sum_{(i,j) \in s} p_{ij}^{(r)} / N \right] \right\}^2. \end{aligned} \quad (3.7)$$

This numerical evaluation is accomplished using a Gibbs sampler; see Malec and Sedransk (1993b) for details.

### 4. Variable Selection

Using data from the 1987 NHIS we select the variables to be included in (2.2) and (2.3) where the binary variable  $Y$  has  $Y=1$  if there has been at least one visit to a physician during the past twelve months. We proceed in two steps by first fitting an individual-level model using (2.1) and (2.2), and then considering the county-level model in (2.3).

Our initial objective is to ascertain the general form of (2.2). We do this by ignoring county variation and estimating  $\beta$  in the "national" model, (4.1). If  $\underline{X}_{ij} = \underline{X}(k)$ , (2.1) and (2.2) are replaced by

$$Pr (Y_{ij} = y_{ij}) = p_k^{y_{ij}} (1-p_k)^{1-y_{ij}}, y_{ij} \in \{0, 1\},$$

and

$$\ln \{p_k/(1-p_k)\} = \underline{X}(k)\underline{\beta}. \quad (4.1)$$

First, we obtain estimates based on the saturated model where the sample proportion of individuals in class  $k$ ,  $\hat{p}_k$ , is used to estimate  $p_k$ . Figure 1 shows the effect of age, race and sex on  $\ln \{\hat{p}_k / (1-\hat{p}_k)\}$ .

The variation in log odds in Figure 1 corresponds to an expected pattern. First, for a given sex and age, the probability of a physician visit is generally larger for whites than for nonwhites. Second, the general patterns are similar for both races for a given sex. For males, the probability of a physician visit decreases steadily until about age 22.5, and then increases steadily. (Recall that we are using five year age groups.) For females, physician visits decrease steadily until age 12.5 and then increase to about age 27.5. Physician visits remain roughly constant from 27.5 until 62.5 and then increase steadily.

Due to the complex form seen from Figure 1, various spline models, linear in age, were used. A fixed knot spline can be defined as a linear model (Smith 1979) and, hence, used in (2.2). We include the possibility of a knot at each five year age group. The general model investigated included all possible splines that are linear in age, a race effect, a sex effect, a race by sex interaction and, finally, all interactions between these categorical variables and the linear age splines. The set of possible variables is

- 1) Categorical variables: Intercept, race ( $r$ ), sex ( $s$ ) and race by sex ( $rs$ )
- 2) Linear age splines:  $X_a(k) = \max(0, \text{age}(k)-a)$ ,  $a=0,5,10,\dots,85$  and  $\text{age}(k)$  is the age for individuals in class  $k$ .
- 3) Categorical by age-spline interactions:  $r$  by  $X_a(k)$ ,  $s$  by  $X_a(k)$ ,  $rs$  by  $X_a(k)$ .

To determine a subset of terms to include in (4.1) the SAS forward stepwise logistic regression procedure, PROC LOGISTIC, was used. This procedure selects variables for inclusion and exclusion using a residual chi-squared test. Since the sample size is approximately 120,000 persons, variables possibly having only a small effect may be included in the model. To determine the total number of variables to use in the model a quantity like  $R^2$  was used. Define the deviance  $D^2$  for the model  $M_1$  as  $(\text{Dev}(M_1) - \text{Dev}(M_0)) / (\text{Dev}(M_s) - \text{Dev}(M_0))$ , where  $M_0$  is the null model (with only an intercept term) and  $M_s$  is the saturated model (a parameter is fitted for each age by race by sex group). Note that

$0 \leq D^2 \leq 1$ , and equals  $R^2$  for the linear model. The variables, intercept,  $sX_{15}, \dots, r$ , listed in the table below were included in the model. Adding other variables does not increase the value of  $D^2$  appreciably (note the small contributions of the next best variables,  $rX_0$  and  $sX_{45}$ , to  $D^2$ ).

Variable	Intercept	$sX_{15}$	$X_0$	$X_{15}$	$sX_{25}$	$X_{35}$	$X_{25}$	$r$	$rX_0$	$sX_{45}$
Cumulative $D^2$	0.00	17.17	22.09	58.88	75.02	87.41	91.55	94.41	95.37	95.70

To check the fit of the model, partial residuals were plotted. Corresponding to each observation there is the residual,  $r_{ij} = (y_{ij} - \hat{p}_{ij}) / \hat{p}_{ij} (1 - \hat{p}_{ij})$ , which is then averaged over subsets of interest. A typical residual plot has, for a given domain (e.g., sex by race),  $r_{ij}$  averaged over all individuals of a given age plotted against age. The particular form of the residual is used because it will estimate a missing term in the logistic model (see Fienberg and Gong's comment to Landwehr, Pregibon and Shoemaker 1984). The residual plot in Figure 2 indicates that the eight variable model provides a good fit to the data. The one large remaining residual (for Black males, aged 85+) corresponds to an estimate based on a very small sample size.

The second step in the data analysis is to identify county-level covariates that affect an individual's probability of visiting a doctor, after having removed the effects due to the individual level covariates. To do this, we combined the individual level and county level models in (2.2) and (2.3) but set  $\Gamma=0$ . Then

$$\ln\{p_{ij}/(1-p_{ij})\} = X_{ij}^t G_i \boldsymbol{\eta} . \quad (4.2)$$

To reduce the scale of this investigation we consider only the eight individual level variables identified earlier. As indicated by (4.2), we allow main effects of county-level variables and interactions of these county covariates with the individual-level variables. The collection of county covariates that we considered are ones included in the Area Resource File or county mortality file, and thought by subject-matter specialists to be relevant. We have also included county variables related to the formation of the NHIS strata. The procedure we used was to force the eight individual-level variables into the model, and let the SAS stepwise logistic regression procedure add variables. (We have also used graphical methods as described in Malec and Sedransk 1993a and Malec, Sedransk, and Tompkins 1993.) We found no county-level covariates that increased  $D^2$  appreciably. However, there is still considerable county-to-county variation to be captured by (2.3) with  $G_i = I$ . For other dependent variables (e.g., health status), county-level covariates play a more significant role.

## 5. Comparison of Alternative Estimators

In this section we use data from the 1987 NHIS to compare the Bayes estimates with the standard alternatives, synthetic and design-based estimates.

For the largest states, the conventional design-based estimates should have relatively small variances, and there should be good agreement between them and estimates based on (3.2). In Figure 3 we plot, for each state and type of estimator (design-based, Bayes, synthetic), the estimated percent of the state population who visited a physician against state sample size. The Bayes estimates (based on a normal approximation to the posterior distribution) are close to the design-based estimates for the largest states, as one would hope. For the same states, the synthetic estimates are always further from the design-based estimates than are the Bayes estimates. As the state sample sizes become smaller the design-based estimates become increasingly unreliable, and the Bayes estimates look less like the design-based estimates, and more like the synthetic estimates. We have also used this same model to produce state estimates of the percent visiting a physician for subpopulations such as persons 65+, non-whites and females. These estimates exhibit the versatility of Bayes estimates; the between county variability, based on  $\Gamma$ , is different for these three cases, leading to different amounts of "gaining of strength". See Malec and Sedransk (1993b) for details and estimates.

Corresponding to Figure 3, Figure 4 is a plot for the 51 areas of posterior standard deviations vs. state sample sizes where we consider both the hierarchical Bayes (formulas (2.1) - (2.4)) and "synthetic"

estimates. For the states with smaller populations, the standard deviations based on the hierarchical Bayes model more properly account for the uncertainty associated with inference about  $\theta$ .

## 6. Evaluation

We have investigated whether the conventional sample weights are informative. Figure 5 is a partial residual plot similar to Figure 2. (For this analysis,

$r_{ij} = \{y_{ij} - E(p_{ij} | \mathcal{Y}_s)\} [E(p_{ij} | \mathcal{Y}_s) \{1 - E(p_{ij} | \mathcal{Y}_s)\}]^{-1}$ .) The ordinate of each point is the average residual for all individuals having a sample weight within the range centered at the corresponding abscissa. There is no evidence that the model should include the sample weight as a covariate.

Since the frequency of persons who visit a physician is not available for the entire NHIS population, it is not possible to compare the small area estimates with the true state values. However, by removing a portion of the sample, cross-validation methods can be used to assess how well the model and estimation procedure predict the part of the sample that has been deleted.

The cross-validation procedure that we plan to use is described below. Define the set of sampled elements that are set aside as "A". Let  $\mathbf{y}_A$  denote the vector of observations that correspond to the elements in A and  $\mathbf{y}_{(A)}$  the remaining sampled elements. Also,  $\mathbf{Y}_A$  is the random variable (with observed value  $\mathbf{y}_A$ ) that represents the removed elements. The predictive distribution,  $f(\mathbf{Y}_A | \mathbf{y}_{(A)})$ , can be used to make comparisons between the observed data,  $\mathbf{y}_A$ , and the values of  $\mathbf{Y}_A$  predicted from the model. Specific functions comparing  $\mathbf{Y}_A$  and  $\mathbf{y}_A$ , denoted  $g(\mathbf{Y}_A | \mathbf{y}_A)$ , can be defined to evaluate features of the predictions. (See Gelfand, Dey and Chang 1991 for a general review of Bayesian model assessment.)

We shall remove sets of sample elements in ways that permit us to see if our model captures the most important features of the NHIS data. Our evaluation will be based on how well the model predicts the deleted sample,

$$\theta_{AU} = \frac{\sum_{i \in U} \sum_{k=1}^K \sum_{j \in A_{ik}} y_{ijk}}{\sum_{i \in U} \sum_{k=1}^K n_{Ai}(k)}$$

where the first sum is over all counties in state "U",  $A_{ik}$  denotes the set of deleted individuals in demographic group k and county i, and  $n_{Ai}(k)$  is the size of  $A_{ik}$ . Two choices for the error in prediction are

$$g_{1U}(\mathbf{Y}_A, \mathbf{y}_A) = (\theta_{AU} - E(\theta_{AU} | \mathbf{y}_{(A)}))^2$$

and

$$g_{2U}(\mathbf{Y}_A, \mathbf{y}_A) = \left( \frac{\theta_{AU} - E(\theta_{AU} | \mathbf{y}_{(A)})}{E(\theta_{AU} | \mathbf{y}_{(A)})} \right)^2.$$

To evaluate how well the model can predict the error of the estimate one may use

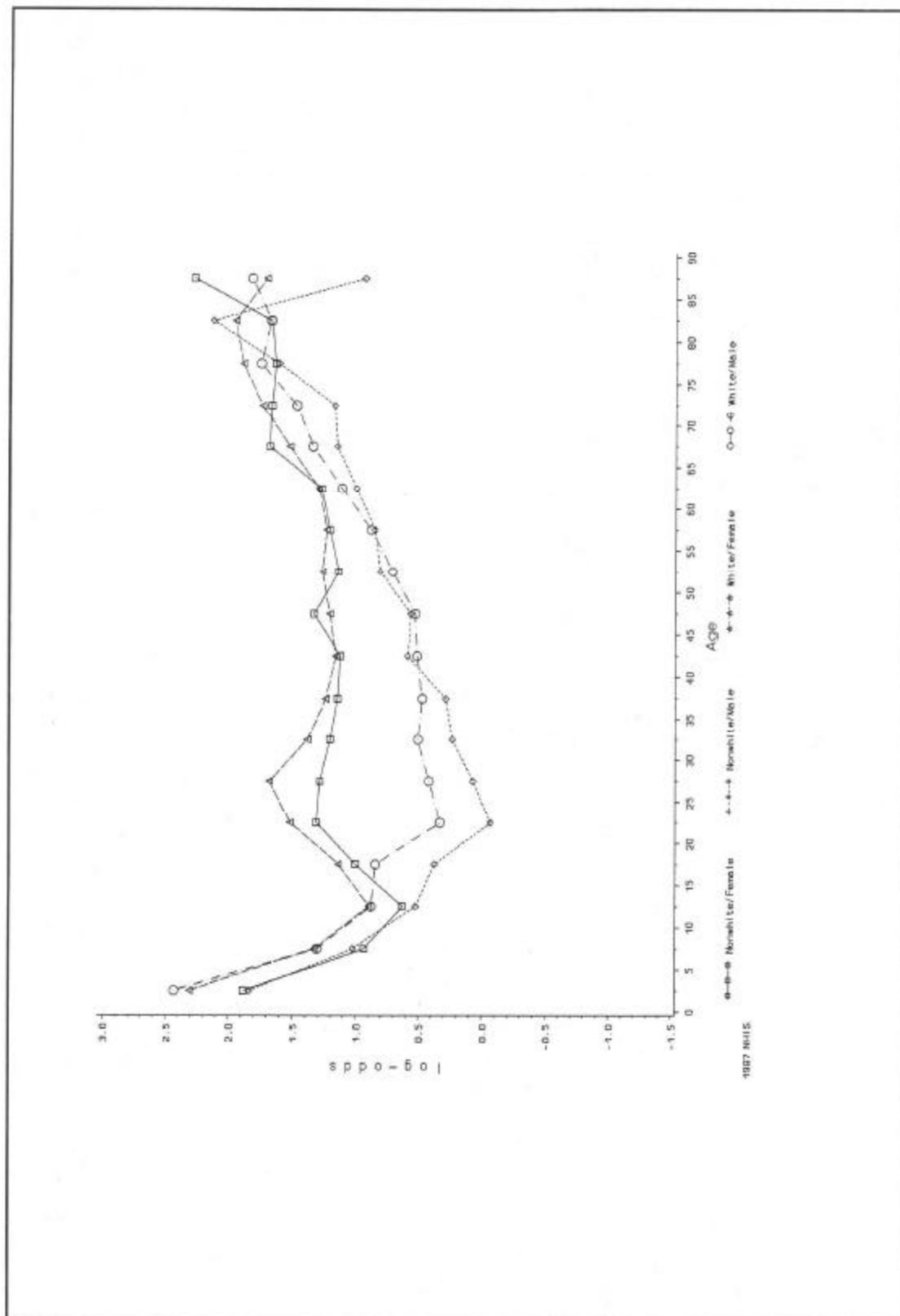
$$g_{3U}(Y_A, Y_A) = \frac{g_{1U}(Y_A, Y_A) - E(g_{1U}(Y_A, Y_A) | Y_{(A)})}{E(g_{1U}(Y_A, Y_A) | Y_{(A)})}.$$

Numerical results from this cross-validation will appear in a forthcoming report.

## References

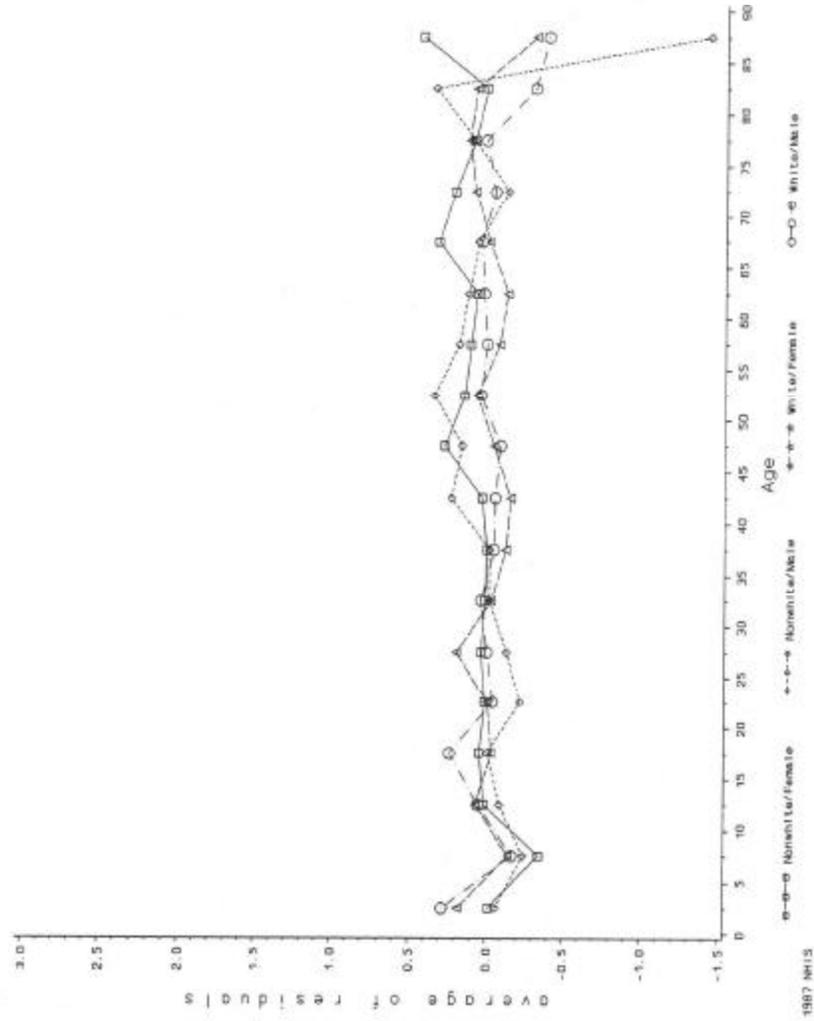
- Dempster, A.P., and Tomberlin, T.J. (1980), "The Analysis of Census Undercount From a Postenumeration Survey," Proceedings of the Conference on Census Undercount, Arlington, VA, 88-94.
- Gatsonis, C., Normand, S-L., Liu, C. and Morris, C. (1993), "Geographic Variation of Procedure Utilization: A Hierarchical Model Approach," *Medical Care*, 31, YS54-YS59.
- Gelfand, A.E., Dey, D.K. and Chang, H. (1991), "Model Determination Using Predictive Distributions with implementations via Sampling-Based Methods". In Bayesian Statistics 4, eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Clarendon Press.
- Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association*, 79, 61-83.
- Malec, D., and Sedransk, J. (1985), "Bayesian Methodology for Predictive Inference for Finite Population Parameters in Multistage Cluster Sampling," *Journal of the American Statistical Association*, 80, 897-902.
- Malec, D., and Sedransk, J. (1993a), "Bayesian Predictive Inference for Units with Small Sample Sizes: The Case of Binary Random Variables," *Medical Care*, 31, YS66-YS70.
- Malec, D., and Sedransk, J. (1993b), "Small Area Inference for Binary Variables in the National Health Interview Survey," Technical Report.
- 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. Springer Verlag.
- Massey, J.T., Moore, T.F., Parsons V.L., and Tadros, W. (1989), "Design and Estimation for the National Health Interview Survey, 1985-94," National Center for Health Statistics. *Vital and Health Statistics*, 2:110.
- Smith, P.L. (1979), "Splines As a Useful and Convenient Statistical Tool," *The American Statistician*, 33, 57-62.
- U.S. Department of Health and Human Services (1989), "The Area Resource File (ARF) System". ODAM Report No. 7-89.
- Wong, G.Y., and Mason, W.M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," *Journal of the American Statistical Association*, 80, 513-524.

Figure 1: Relationship of log odds of a sample proportion and age for each race by sex group.



NOTE: The binary variable is the presence or absence of at least one visit to a physician within the past year.

Figure 2: Relationship for each sex x race group of average residual and age using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all members of a given sex by race by age group.

Figure 3: Estimated percent of population in a state who visited a physician in the past year plotted against state sample size: Hierarchical Bayes, synthetic and design-based estimates.

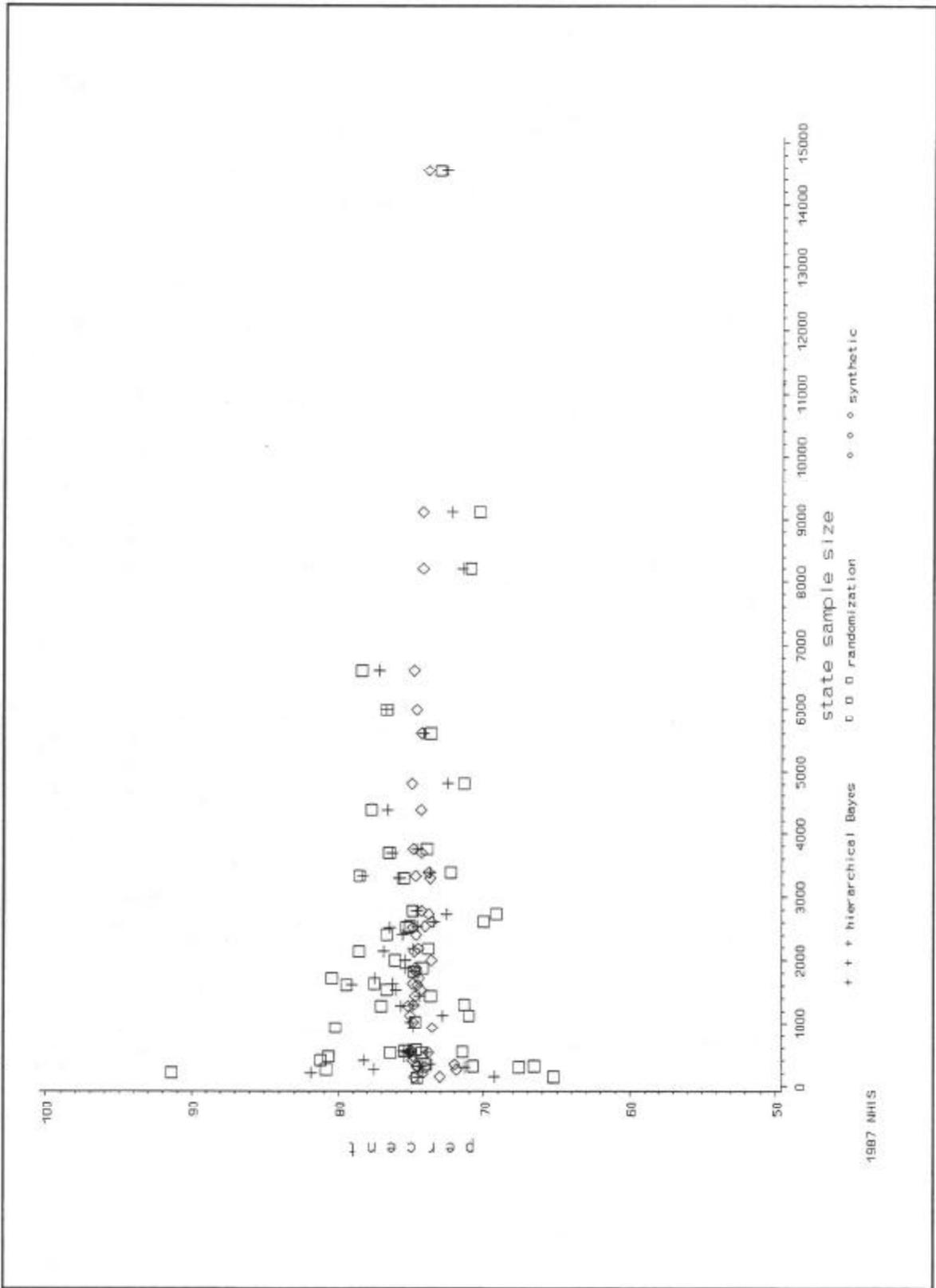


Figure 4: Posterior standard deviations plotted against state sample sizes: Hierarchical Bayes and synthetic models.

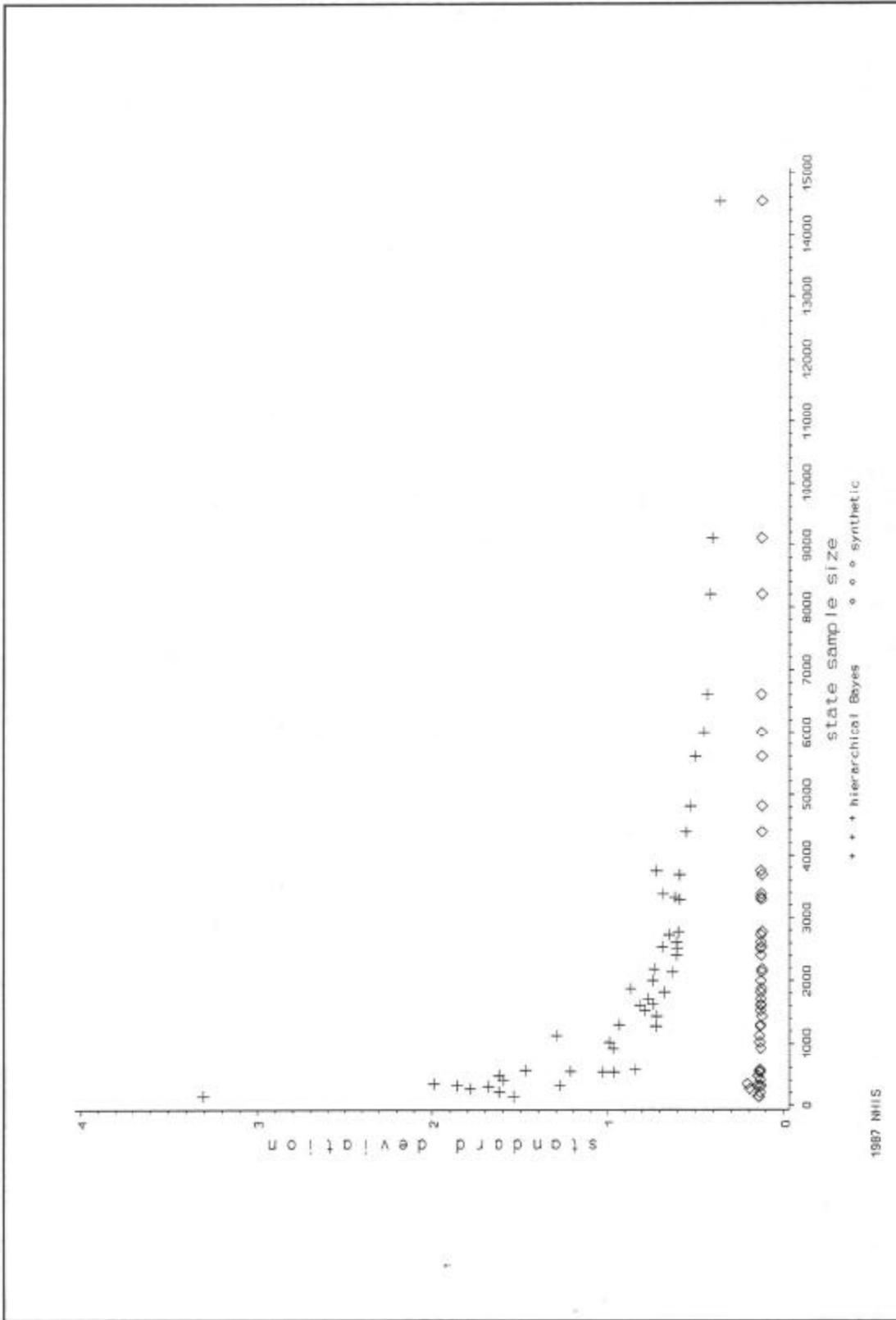
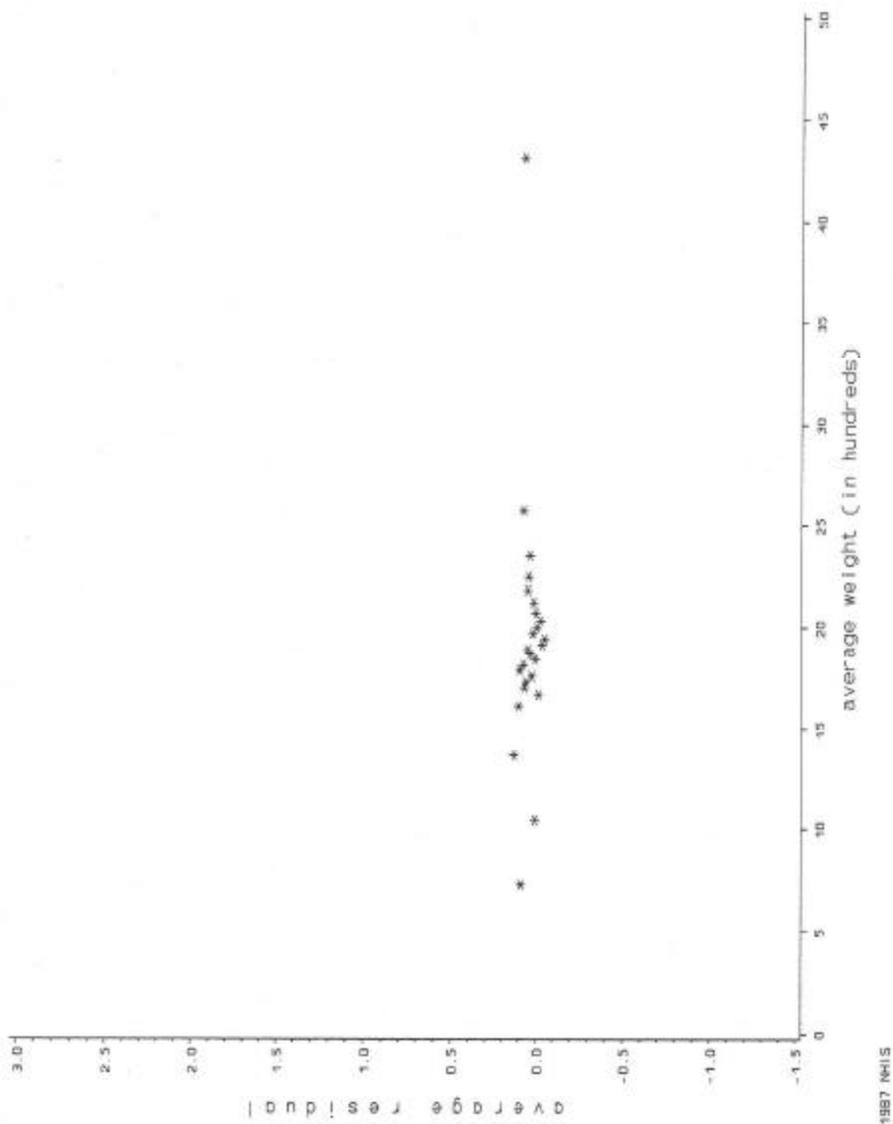


Figure 5: Relationship of average residual and sample weight using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all individuals having a sample weight within a narrow range.

THE ROLE OF DESIGN BASED VARIANCES AND COVARIANCES  
IN SMALL DOMAIN ESTIMATION

Robert E. Fay<sup>1</sup>  
U.S. Bureau of the Census

1. Introduction

Two recent reviews provide the context for this paper. The Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology (Schaible and Gonzalez 1993) surveyed applications of indirect estimators in official U.S. government use. The first chapter of their report noted the predominance of direct estimators in federal statistics. In other words, official estimates are almost always "direct," through exclusive or almost exclusive use of data from both the time and domain of interest. Indirect estimators, relying heavily on data from either other domains or times (or both), are the exception in federal statistics. The report enumerates and discusses indirect estimators in current federal use. (That is, the report considered only those applications published as official estimates, not including methodological tests and discontinued series. Generalizations of survey variance estimates, for example, those often included in source and reliability statements at the end of Census Bureau reports, were also not included.)

Although infrequently employed in federal practice, indirect estimation generally reflects an attempt to address a need for estimates that cannot be reliably produced directly given constraints on resources. The concluding chapter of the report urged caution in the use of indirect methods and eschewed advocacy of them as a general purpose and easily developed solution.

Ghosh and Rao (1994) reviewed the statistical methodology underlying several types of indirect estimators. Their review included demographic and other methods specific to postcensal population estimation; synthetic, composite and related estimators for domain characteristics; and empirical best linear unbiased predictors (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). This paper employs their review as a point of departure for comparisons of existing theory to practice.

Several small domain applications that have appeared in the literature share enough common features to be studied as a group. One class of applications, which represents the scope of this paper, combines information from survey estimates at the domain

---

<sup>1</sup> This article represents results of research undertaken by a staff member of the Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the U.S. Bureau of the Census.

level with domain-level characteristics available from independent sources. Examples include estimates of 1970 census income for small places (Fay and Herriot 1979), estimates of 1980 census undercount (Erickson and Kadane 1985, Cressie 1992), estimates of 1990 census undercount produced by the U.S. Census Bureau in 1991 (subsequently revised under a different methodology), and estimates of median family income by state (Fay, Nelson, and Litow 1993).

In some of these applications, independent data provide a basis for evaluating the methodology. For example, in estimating median family income by state, the decennial census figures serve as a gold standard by which to judge the performance of the resulting small domain estimates. Although this comparison is available only every 10 years, the empirical results support the application. As a second example, the relatively small number of available special censuses taken after the 1970 census also corroborated the application to 1970 census income for small places. In other cases, however, including the analysis of census undercount, there is no gold standard by which to evaluate the resulting estimates. Consequently, the validity of the application of the underlying theory for both the properties of the resulting estimates and the measurement of their reliability is of considerable importance.

Comparison of these and other applications to the available theory generally shows that the explicit theoretical conditions are not completely satisfied, although to varying degrees. Consequently, each application implicitly requires that the departures from the theory do not pose serious consequences. As the title of this paper suggests, the theoretical results typically assume that the sampling errors of the small domain estimates are known, whereas in practice they are frequently estimated from the data, either directly or through a model to generalize the variances.

Section 2 reviews much of the existing theory for the class of estimators under discussion. Section 3 then compares the applications just mentioned to the requirements of the theoretical formulations to note implicit extensions of the theory that, for the most part, still lack a theoretical foundation. Section 4 reports the results of simple Monte Carlo studies to assess evidence in some of these areas. Although mathematical proof is preferable to computer demonstration, the empirical results present useful evidence on the significance of various issues arising from the practical application of these procedures and suggest directions for new research.

## 2. Theoretical Results for a Class of Small Area Estimators

As noted in the previous section, Ghosh and Rao (1994) reviewed several general small area approaches. The class of

models of interest to this paper employs auxiliary data  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ , which are assumed measured without sampling error. In their notation, the parameters of interest,  $\theta_i$ , are assumed to be related to the  $\mathbf{x}_i$  by

$$\theta_i = \mathbf{x}_i \beta + v_i z_i, \quad i = 1, \dots, m, \quad (2.1)$$

Frequently, the model takes the simpler form:

$$\theta_i = \mathbf{x}_i \beta + v_i, \quad i = 1, \dots, m, \quad (2.2)$$

where  $\beta$  is a vector of regression parameters, and the  $v_i$  are independent, identically distributed (iid) random variables with:

$$E(v_i) = 0, \quad V(v_i) = \sigma_v^2. \quad (2.3)$$

In (2.1), the  $z_i$ 's are known positive constants. Ghosh and Rao (1994) develop the theory in the general form (2.1). Results specific to the simpler model (2.2) are offered here because the formulas are more accessible.

The  $\theta_i$  in (2.1) and (2.2) represent the parameters of interest for the small areas, such as local area per capita income, the ratio of correction population to census population, the number of employed, etc. The model reflects a possible lack of fit between the regression  $\mathbf{x}_i \beta$  and the actual value through random effect terms,  $v_i$ .

In this class of models, direct estimates,  $\hat{\theta}_i$  are available at the domain level with

$$\hat{\theta}_i = \theta_i + e_i, \quad (2.4)$$

where the  $e_i$  represent sampling errors with

$$E(e_i | \theta) = 0, \quad V(e_i | \theta) = \psi_i. \quad (2.5)$$

(In this section, the sampling errors are also assumed independent, but extensions have reflected correlated sampling errors.) In other words, the  $\hat{\theta}_i$  are design-unbiased estimators. Ghosh and Rao comment that these conditions may be quite restrictive. For example, the estimators may not be unbiased, as in the case of undercount adjustment. In addition, the sampling variances  $\psi_i$  may not be known.

The combined model, using (2.2) and (2.4), is

$$\hat{\theta}_i = \mathbf{x}_i \beta + v_i + e_i. \quad (2.6)$$

As Ghosh and Rao note, (2.6), which is a linear combination of fixed and random effects, is a special case of the general mixed linear model.

Ghosh and Rao (1994) discuss the estimation of (2.6) from the perspectives of EBLUP, EB, and HB; this paper will primarily focus on the EBLUP formulation. They cite Henderson (1950) as the originator of best linear unbiased predictors (BLUP) for models such as (2.6), when the variance components are known. Ghosh and Rao express the BLUP of  $\theta_i$  as

$$\hat{\theta}_i^H = \gamma_i \hat{\theta}_i + (1 - \gamma_i) \mathbf{x}_i \hat{\beta}, \quad (2.7)$$

where

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \hat{\theta} \quad (2.8)$$

is the BLUE of  $\beta$ ,  $\mathbf{V}$  is the diagonal matrix with elements  $\sigma_v^2 + \psi_i$ , and

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \psi_i}. \quad (2.9)$$

When the variance components are known, the mean square error of (2.7) under model (2.6) is

$$E(\hat{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2), \quad (2.10)$$

where

$$g_{1i} = Y_i \psi_i, \quad (2.11)$$

and

$$g_{2i}(\sigma_v^2) = (1 - Y_i)^2 \mathbf{x}_i (\mathbf{x}' \mathbf{V}^{-1} \mathbf{x})^{-1} \mathbf{x}_i' . \quad (2.12)$$

Because  $\sigma_v^2$  is typically unknown, a two-stage estimator,  $\hat{\theta}_1^H$ , arises by first estimating  $\sigma_v^2$  from the data and then using it to obtain (2.7). Ghosh and Rao reference several options for estimating  $\sigma_v^2$ . A simple moment estimator  $\hat{\sigma}_{v(1)}^2 = \max(\sigma_{v(1)}^2, 0)$ , where

$$\hat{\sigma}_{v(1)}^2 = (n - p)^{-1} \left[ (\hat{\theta} - \mathbf{x}\beta^*)' (\hat{\theta} - \mathbf{x}\beta^*) - \sum_i \psi_i \{1 - \mathbf{x}_i (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}_i'\} \right], \quad (2.13)$$

and

$$\beta^* = (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' \hat{\theta}$$

is the ordinary least squares estimator of  $\beta$ , has the advantage of not requiring iteration.

The remaining methods to be considered here each require iteration, unless the sampling errors  $\psi_i$  are equal. For a given trial estimate of  $\sigma_v^2$ ,  $\beta$  is estimated through (2.8) at each cycle for each of the methods.

Fay and Herriot (1979) used an estimator  $\hat{\sigma}_{v(2)}$ , based on the method of moments as the solution to the equation

$$(\hat{\theta}_i - \mathbf{x}\beta)' \mathbf{V}^{-1} (\hat{\theta}_i - \mathbf{x}\beta) = n - p \quad (2.14)$$

or 0 in the case that no solution exists. If the sampling errors  $\psi_i$  are equal, then (2.13) and (2.14) have the same solution.

A third alternative, maximum likelihood,  $\sigma_{v(3)}$ , maximizes

$$L(\beta, \sigma_v^2) = -1/2 \log(|V|) - 1/2 (\theta - X\beta)' V^{-1} (\theta - X\beta), \quad (2.15)$$

which is the log-likelihood up to a constant.

Cressie (1992) suggested the application of another estimator,  $\sigma_{v(4)}$ , based on restricted maximum-likelihood (REML), which maximizes the adjusted likelihood,

$$L^*(\beta, \sigma_v^2) = -1/2 \log(|V|) - 1/2 (\log(|X'V^{-1}X|) - 1/2 (\theta - X\beta)' V^{-1} (\theta - X\beta)), \quad (2.16)$$

where terms not involving the parameters have been dropped. Cressie (1992) further describes this procedure, which was originally developed by Patterson and Thompson (1971, 1974). In short, however, the procedure examines the likelihood of the residuals from the regression. When all the sampling errors  $\psi_i$  are equal, then (2.13), (2.14), and (2.16) have the same solution, while (2.15) yields a generally smaller estimate of  $\sigma_v^2$ .

Under normality of the error terms, Kackar and Harville (1984) showed that

$$E(\hat{\theta}_1^H - \theta_1)^2 = E(\hat{\theta}_1^H - \theta_1)^2 + E(\hat{\theta}_1^H - \hat{\theta}_1^H)^2,$$

where, for large  $m$ , the second term may be approximated by

$$g_{31}(\sigma_v) = \psi_1^2 (\sigma_v^2 + \psi_1)^{-3} \bar{V}(\sigma_v^2),$$

where  $\bar{V}(\sigma_v^2)$  is the asymptotic variance of  $\sigma_v^2$ .

Prasad and Rao showed that an approximately unbiased estimator of the mean square error of the EBLUP estimator is

$$mse(\hat{\theta}_1^H) = g_{11}(\sigma_v^2) + g_{21}(\sigma_v^2) + 2g_{31}(\sigma_v^2) \quad (2.17)$$

with bias of order lower than  $m^{-1}$ .

The MSE estimators studied in Section 4 share (2.17) but differ in the approach to estimate  $\bar{V}(\sigma_v^2)$ . Section 4 describes these differences.

### 3. Assumptions Made by Some Previous Applications of EBLUP

Section 2, and the more complete review in Ghosh and Rao (1994), detail the assumptions of the available theory for EBLUP. This section briefly reviews potential discrepancies between the theory and some previously published applications.

One feature is common to all of the applications discussed here and can be assumed to occur almost universally for sample surveys, namely, that the sampling variances,  $\psi_i$ , are estimates rather than known values. The following discussion notes the consequent adaptations, which range from direct use of estimated variances to variance generalization.

Fay and Herriot (1979) described a large-scale implementation of EBLUP/empirical Bayes estimation to estimate per capita income in 1969 for small places and minor civil divisions with population below 1000 persons. The sample estimates,  $\hat{\theta}_i$ , were based on the long form sample of the 1980 census. Predictors included the county average PCI, 100% data from the census on housing value, and reported income from IRS returns. Because of computing constraints at the time, the authors refrained from any recalculation of the census sampling variances but instead employed the available variance generalization. The generalization was a simple national model without any allowance for geographic variation. Since the generalization yielded a linear relationship between  $\psi_i$  and  $\theta_i^2$ , a logarithmic transformation of  $\hat{\theta}_i$  gave a closer fit of the application to the theory. They employed (2.14) to estimate  $\sigma_v^2$ . Generally, the compositing, (2.7), drew on both the sample estimates and the regression in approximately equal amounts, rather than relying almost exclusively on one of the two. The authors employed (2.2) but observed some evidence of variation in  $\sigma_v^2$  by size of place. The evidence suggested (2.1) with  $z_i$  decreasing with increasing size,  $n_i$ , although at a rate closer to  $n_i^{-1/4}$  than  $n_i^{-1/2}$ . The authors did not attempt MSE estimation, but presented some limited empirical evidence from special censuses favoring the EBLUP approach.

Application of EBLUP to sample estimates of decennial census undercounts has been controversial, and the review here will simply focus on assumptions incorporated in the implementations rather than systematically evaluating the merits of the work on this subject. As Ghosh and Rao (1994) comment in passing, survey estimates of undercount in both 1980 and 1990 have been subject to substantial sources of bias, and the existing theory does not provide a clear measure of how EBLUP behaves under such conditions. Furthermore, gains from EBLUP and estimators of MSE have figured

prominently in the undercount debate, since the 1980 PES estimates at the state level and the 1990 PEP estimates based on the original 1392 strata have such high sampling variability as to preclude adjustment without EBLUP or other smoothing. The estimators placed high weight on the regression and little on the direct estimates.

In both 1980 and 1990, estimates of  $\psi_i$  have appeared to depend on  $\hat{\theta}_i$ . The published 1980 analysis used the estimated variances in spite of this departure from the model. The 1991 analysis of the 1990 PES applied a variance generalization. Although opinions have been offered on the subject, a systematic analysis of the effect of the generalization on the 1991 estimates remains to be done. Furthermore, the 1991 smoothing was multivariate and employed large covariance matrices, formed from the generalized variances and directly estimated correlations. Fay (1992) showed through stratified bootstrap samples that this approach induced substantial additional variability not reflected in the MSE's computed by the Census Bureau.

The 1980 PES estimates were subject to substantial amounts of missing data, yet no estimates of missing data variance are available, and the author is unaware of systematic analysis showing what possible effect this factor might have had on the 1980 analysis.

Ericksen and Kadane (1985) and the 1991 EBLUP for the 1990 PES both employed (2.2), whereas Cressie (1992) reanalyzed 1980 estimates with  $z_i \cdot n^{-1/2}$ . Although Cressie argued for this choice on intuitive grounds, empirical evidence on this question is limited and virtually impossible to obtain from the undercount estimates themselves. The 1990 application employed (2.2); yet the sample estimates suggested that it failed to hold because  $\sigma_v^2$  appears much larger in minority poststrata than elsewhere.

The U.S. Census Bureau has employed an EBLUP procedure to estimate median family income for 4-person families by state from the Current Population Survey (Fay, Nelson, and Litow 1993). The model can be calibrated against census values every 10 years. These calibrations have favored continued use of (2.2) at the state level, distinctly rejecting proposals such as  $z_i \cdot n^{-1/2}$  (Cressie 1992) in this application. The authors account for different approaches to estimating  $\psi_i$  and  $\sigma_v^2$  over the evolution of the model. Over time, more emphasis has been placed on direct estimates.

In short, 1) each of these applications has rested on implicit extensions of the existing theory, 2) some empirical evidence suggests that these procedures can be useful under some conditions, but 3) a more systematic approach to assessing effects of uncertainty for EBLUP is still needed. The next section does not fully meet this need, but it does suggest the value of large scale Monte Carlo simulation as a productive approach to some of these questions.

## 4. Monte Carlo Evaluation

### 4.1 Basic Design of the Study

As noted earlier, the derivation of the estimators of mean square error rest on expectations taken both over repetitions of the sample and over the random effects. The more usual perspective of finite population sampling considers the population as fixed but unknown. In order to bridge the consequences of these two points of view, this study generated several finite populations,  $\theta$ , and compared the properties of the mean square error estimates for each resulting population. Although the expressions in Section 2 focused on errors for individual components, (2.17), this section studies the accuracy of the estimated sum over domains of squared errors, much as the literature on the James-Stein estimator. The actual MSE's are compared to (2.17) summed over  $i$ .

Two values of  $m$ , 20 and 50, offer some indication of the effect of number of domains on the estimators. The first offers an approximate lower boundary on the range of usual application, while the second illustrates the effect of somewhat larger  $m$ . The primary emphasis will be on 50. Obviously, results for larger numbers of domains, such as 200, would also be desirable.

The section reports results that share the following common elements:

- 1) A set of population values for the domains,  $\theta$ , is drawn from some distribution. For normal applications, for example, the domain means are selected.
- 2) Samples,  $\hat{\theta}$ , are drawn from the domain population and sampling variances estimated.
- 3) One or more EBLUP's,  $\hat{\theta}^n$ , are constructed.
- 4) Steps 2) and 3) are repeated for a total of 2000 samples from the population defined at step 1).
- 5) The MSE's of the EBLUP's, estimated as the average value of  $(\hat{\theta}_i^n - \theta_i)^2$ , calculated from the 2000 samples and summed over the domains,  $i$ , become the standard for comparison to the corresponding estimated MSE sums of (2.17) over  $i$ .
- 6) The bias and mean square error of the MSE estimators are then derived, and also key frequencies, such as the percent of samples in which the estimated MSE understates the actual MSE by 25 percent.

Steps 1)-6) generate one point in the Monte Carlo study. In other words, each point represents a specific population realized from

the superpopulation, where the performance, over repeated sampling, of each of the EBLUP estimators and MSE estimators is evaluated.

Obviously, the MSE's at step 5) are not entirely free from sampling error themselves, but the relatively large number of samples provides practical justification for this procedure. The results presented in Figures 1-13 show that this procedure produces stable values.

Thus, the perspective is similar to design-based finite population sampling, since the criteria assess the performance for individual over repetitions of the sampling design.

#### 4.2 Results for Normal Populations

For simplicity, four X variables were constructed:

- 1) The grand mean.
- 2) An indicator variable dividing the domains into halves according to domain number,  $i = 1, \dots, m$ . For example, for  $m = 20$ , the variable distinguishes the first 10 from the second.
- 3) A linear term, increasing with the domain number.
- 4) A similar quadratic term.

The sample size,  $n$ , for each domain was fixed at either 10 or 20, and the observations were scaled by  $n^{1/2}$  in order to give the sample means unit variance. Because the analysis is invariant to the true  $\beta$ , these coefficients were set to 0 in generating the Monte Carlo samples.

Eight combinations were studied:

- 1) Use of known sampling variances,  $\psi_i = 1$ , in combination with (2.14) and  $\bar{V}(\sigma_v^2)$  from Prasad and Rao (1990, p. 167, (5.19)):

$$\bar{V}(\sigma_v^2) = 2m^{-1} \left[ \sigma_v^4 + 2\sigma_v^2 \sum \psi_i/m + \sum \psi_i^2/m \right] \quad (4.1)$$

- 2) Use of generalized sampling variances assuming that the  $\psi_i$  are equal to some unknown constant, which is then estimated as the average of the sample estimates of  $\psi_i$ . The remaining estimation is done as in 1).
- 3) MLE using known sampling variances,  $\psi_i = 1$ , and:

$$\bar{V}(\sigma_v^2) = 2 \left[ \sum (\sigma_v^2 + \psi_i)^{-2} \right]^{-1} \quad (4.2)$$

- 4) MLE using the estimated variances and the estimation approach of 3).
- 5) REML using the estimated variances and  $\bar{V}(\sigma_v^2)$  from Cressie (1992, p. 82, (3.22) and p. 85, (4.11)).
- 6) REML using the estimated variances and the more approximate expression (4.2).
- 7) The method of moments estimator, (2.14), and:
 
$$\bar{V}(\sigma_v^2) = 2m \left[ \sum (\sigma_v^2 + \Psi_i)^{-1} \right]^{-2} \quad (4.3)$$
- 8) The simple moment estimator, (2.13), and (4.1).

Cressie's (1992) estimator studied included in 5) is the only one of the group to explicitly incorporate the effect of the regression in estimating  $\bar{V}(\sigma_v^2)$ . All others depend on  $m$  being large compared to  $p$ . In fact, however, differences between 5) and the simpler 6) were extremely modest. Potentially, similar refinements could be incorporated into the other estimators of  $\bar{V}(\sigma_v^2)$ , but their impact is again likely to be small unless  $p$  is a substantial proportion of  $m$ .

Figures 1-13 results for  $m = 50$ ,  $n = 10$ , that is a comparatively large number of domains with comparatively few degrees of freedom in each domain to estimate the variance in each. Of course, no one choice of these values is appropriate to represent the usual situation in most small domain estimation. Comments will follow about the results obtained for  $m = 20$  and for  $n = 20$ .

A series of 28 populations are represented: 4 drawn from  $N(0, \sigma_v^2)$  with  $\sigma_v^2 = .125$ , and 8 each from  $\sigma_v^2 = .25, .5, \text{ and } 1.0$ . Figures 2, 4, 6, 8, 10, and 12 each omit the points for  $\sigma_v^2 = .125$ , which are generally far off the scale; further comments on this point follow.

Figure 1 shows the actual MSE for 2) as a function of  $\sum \theta_i^2$ , which is called the "SS of true deviations" in the figures. Over the entire range studied, the EBLUP improves on the direct sample estimates, but the improvement is most dramatic at the leftmost portion of the range, where  $\sigma_v^2 = .125$ , and the true values  $\theta$  almost fit the regression line. The pluses and x's distinguish between different super population values of  $\sigma_v^2$  used to generate the  $\theta_i$ ,

but quite clearly this distinction is unimportant once the results are conditioned on  $\sum \theta_i^2$ .

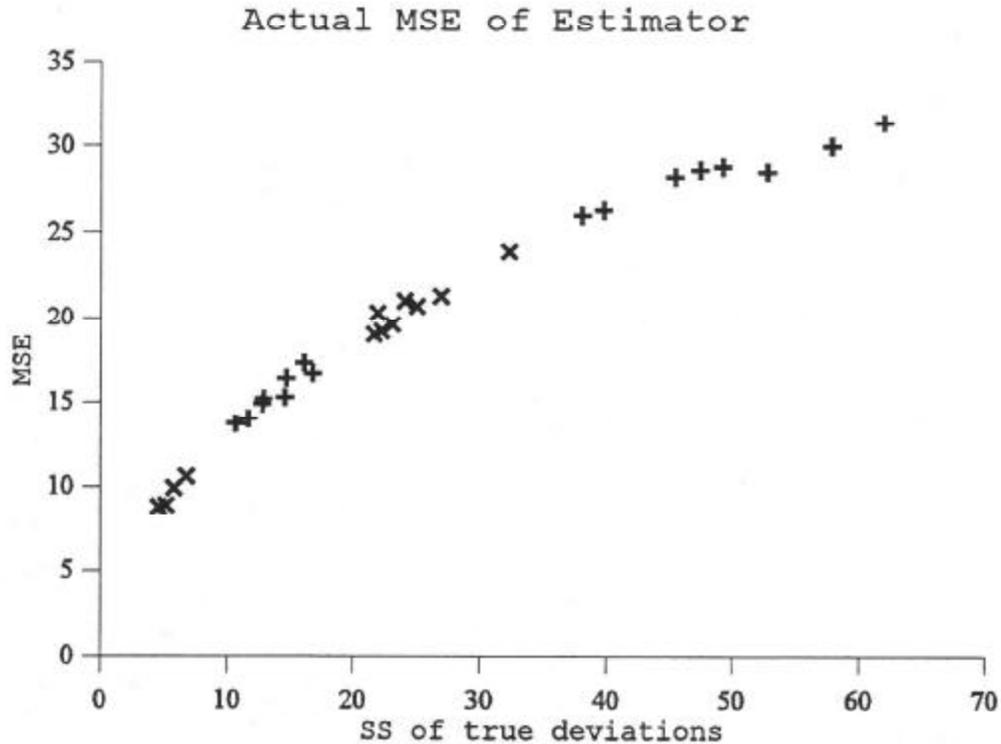
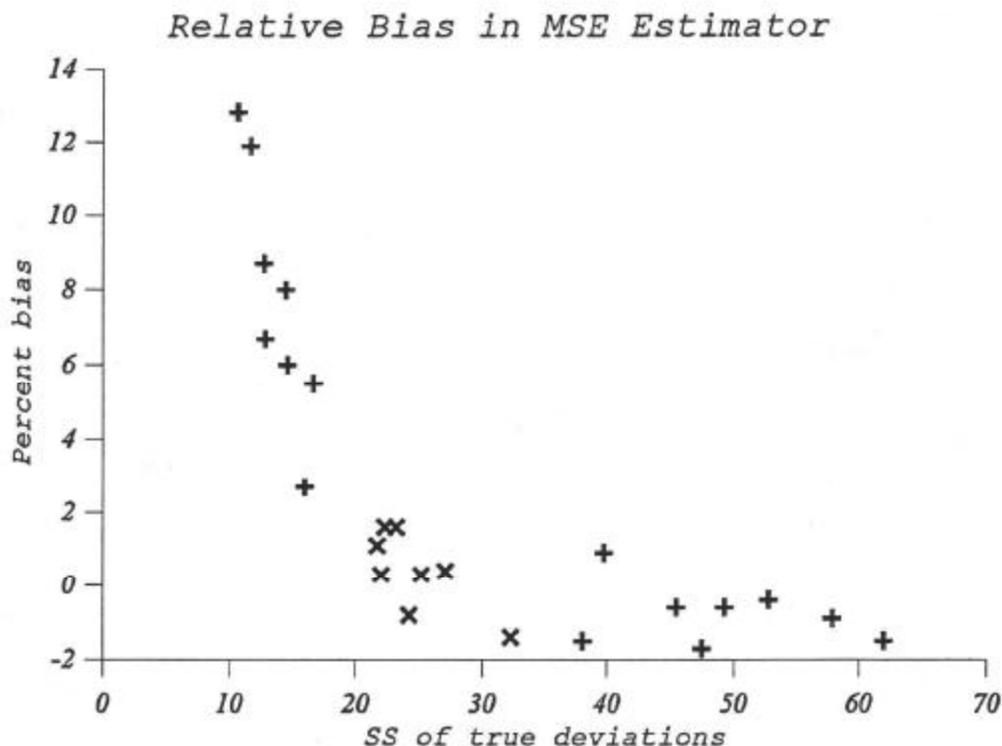


Figure 1 Actual MSE for "unbiased methods," generalized equal variances,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

If the actual MSE's of any of the alternatives were superimposed on Figure 1, there would be substantial overlap. The MSE's and other performance characteristics of 1), with known variances, are virtually identical to 2). The actual MSE's of 3), MLE with known variances, are also almost identical to those in Figure 1. When sampling errors are instead estimated, the actual MSE's are a bit larger: by about 15-30% for REML and 8-20% for the other alternatives when the actual MSE is below 20, and by lesser amounts over the upper end of the range.

Since the MSE of the sample means is 50, Figure 1 includes a broad range of outcomes. At MSE=30, EBLUP yields distinct gains that, nonetheless, many practitioners might choose to forego in favor of the greater simplicity and interpretability of the direct

sample estimates. At MSE=15, the gains from EBLUP may have a substantial impact on the utility of the estimates.



**Figure 2** Percent bias in Prasad-Rao MSE estimator, generalized equal variances, "unbiased methods,"  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 2 reports the relative bias of the MSE estimator for 2) over the range of  $\sum \theta_i^2$ . The leftmost 4 points have been omitted from the graph because the bias increases dramatically, to around 30-40%, in that region. As noted previously, the performance for 1), with known sampling variances, is virtually identical to Figure 2.

From the perspective of bias, the performance of the MSE estimator is quite satisfactory over a large part of the range, but it becomes upwardly biased under conditions where the EBLUP estimator has the most pronounced effect, that is, in the leftmost portion of the range, below MSE = 15 or so.

Figure 3 evaluates the performance of the MSE estimator in a different manner, by showing the proportion of times that the estimated MSE falls below the actual MSE by 25% or more. For example, when the actual MSE=20, the figure reports the percentage of samples in which the estimated MSE is below 15.

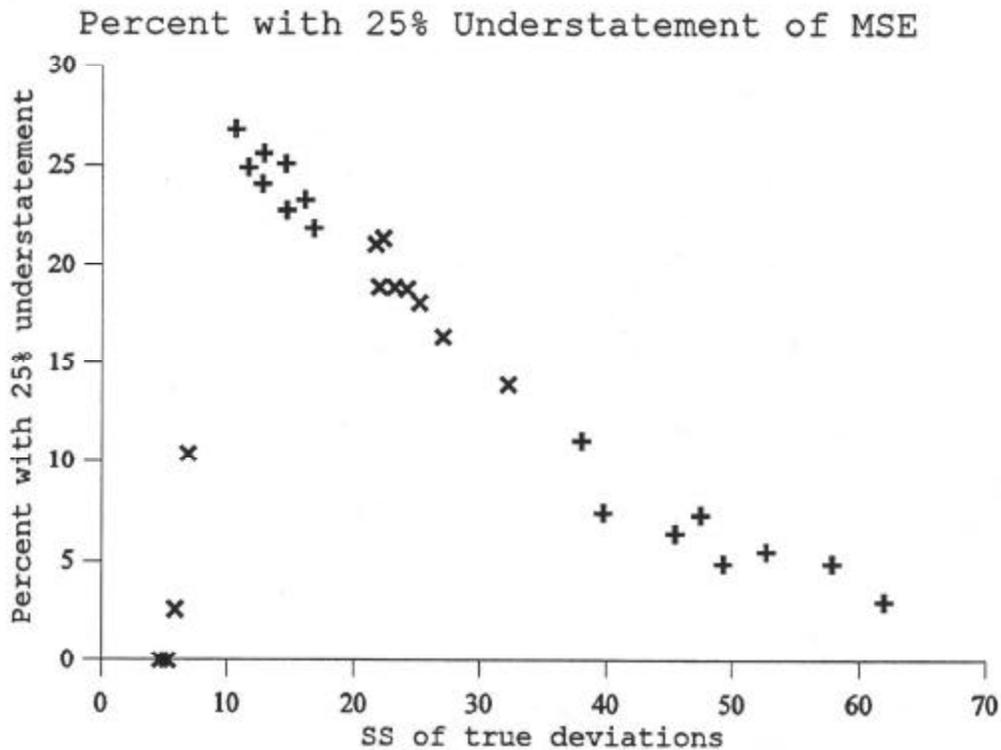


Figure 3 Percent understatement of true MSE by 25 percent or more, generalized equal variances, "unbiased methods,"  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

The findings of Figure 3 are not easily predicted from Figure 2. In spite of the low level of bias in the MSE estimator over the range of MSE=20 and above, the probability that the estimator will substantially understate the actual MSE rises steadily as MSE decreases. Even more striking, however, is the dramatic fall towards 0% at the right of the figure. In fact, in this lower range, the contributions of the more stable components of (2.17), namely its second and third terms, are able to prevent a large understatement regardless of the contribution of the far more erratic first term.

Figure 4 presents comparable results for 3), MLE with known variances. Figure 4 reports a consistent downward bias in the estimated MSE for MLE. This finding agrees with a comparison of REML and MLE by Cressie (1992). Presumably, this downward bias could be even more severe when the ratio of  $p$  to  $m$ , which is 4 to 50 in this case, is larger.

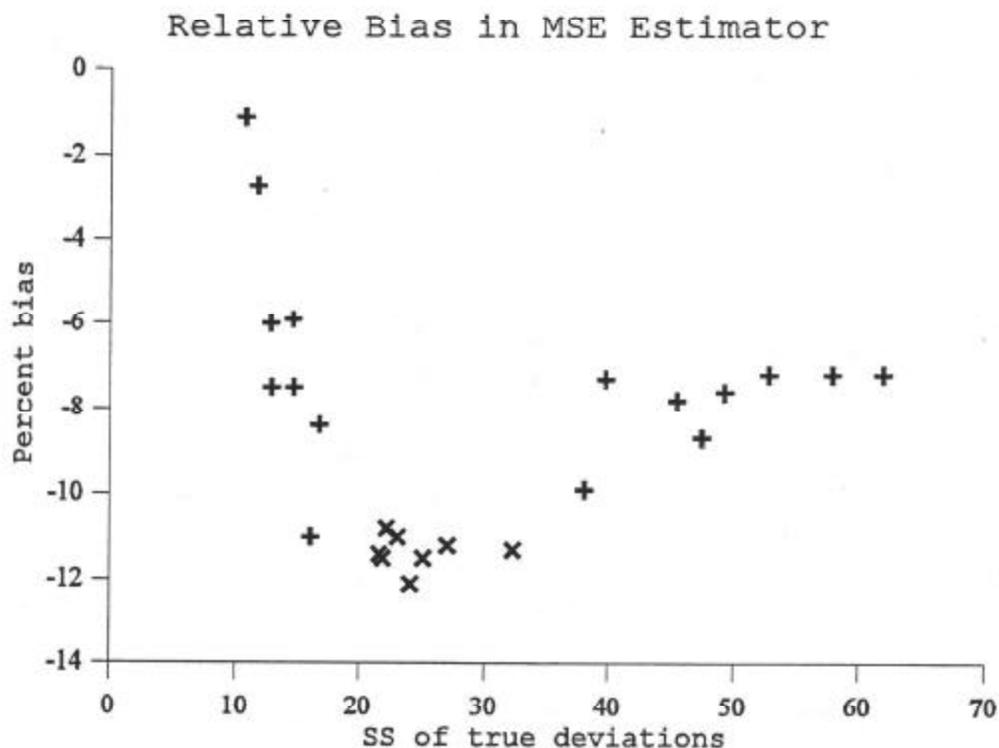


Figure 4 Percent bias in Prasad-Rao MSE estimator, generalized equal variances, MLE,  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

In spite of the general downward bias in the MSE estimate, the bias changes sign and increases up to about 15-30% for the 4 lowest points included in the study.

Figure 5 presents results for MLE analogous to those in Figure 3. Noting the change in scale between the two figures, Figure 5 shows even higher proportions of significant understatement of the MSE over a large proportion of the range. This finding is consistent with the general downward bias exhibited in Figure 4.

As in Figure 3, however, the probability of significant understatement falls off dramatically near 0.

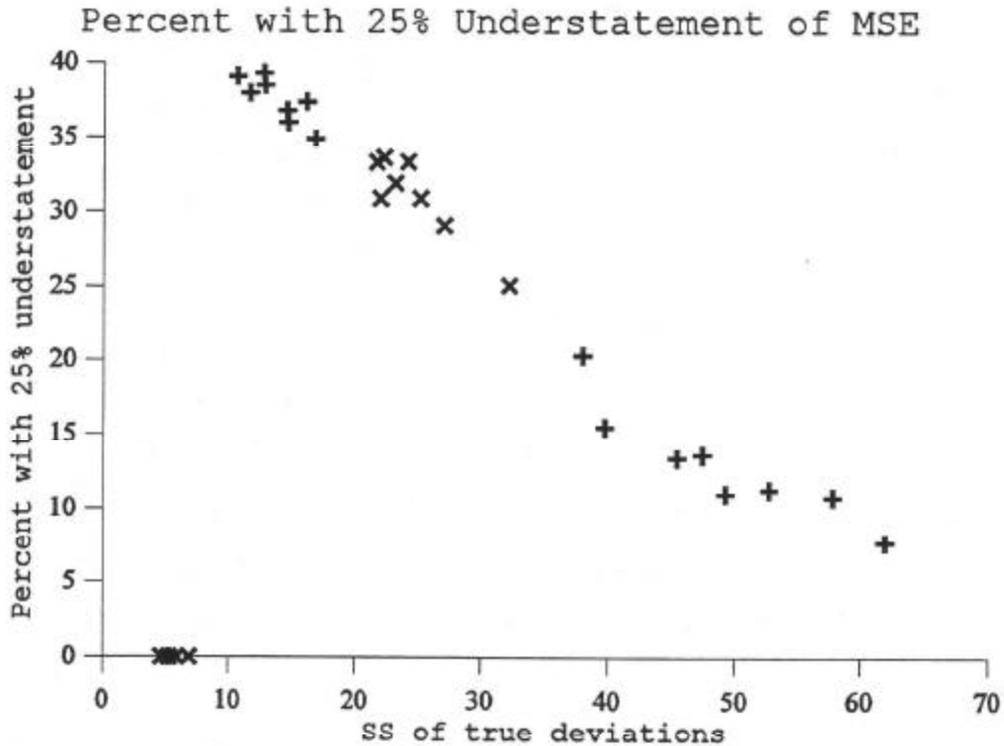


Figure 5 Percent understatement of true MSE by 25 percent or more, generalized equal variances, MLE,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

As noted earlier, shifting from known variances to estimated variances for each of the domains increases the actual MSE of the MLE by about 8-20% for actual MSE's below 20, and somewhat less for larger actual MSE's. Figure 6 reports the performance of the MSE estimator in this instance, as an estimator of the actual, and now larger, MSE. Comparison of Figures 4 and 6 indicates some common features but considerable differences as well. On the right of Figure 6, the downward bias is even more pronounced than in Figure 4. For decreasing MSE, however, the bias crosses 0 earlier than in Figure 4. The bias for the omitted points rises to approximately the same range, that is, about 15-30%.

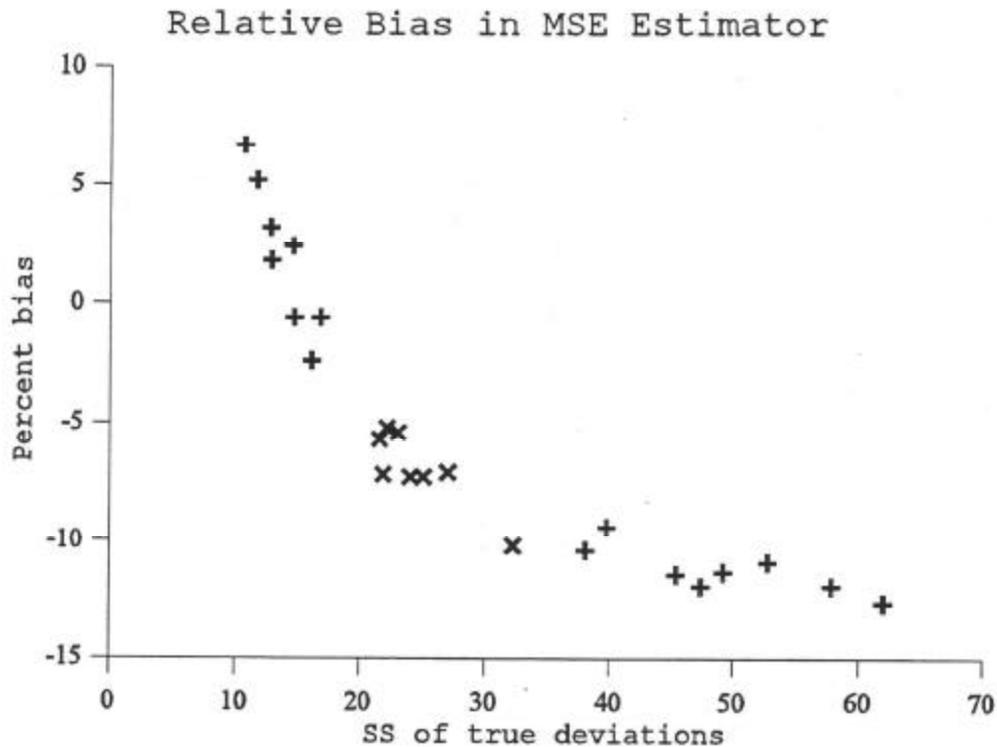
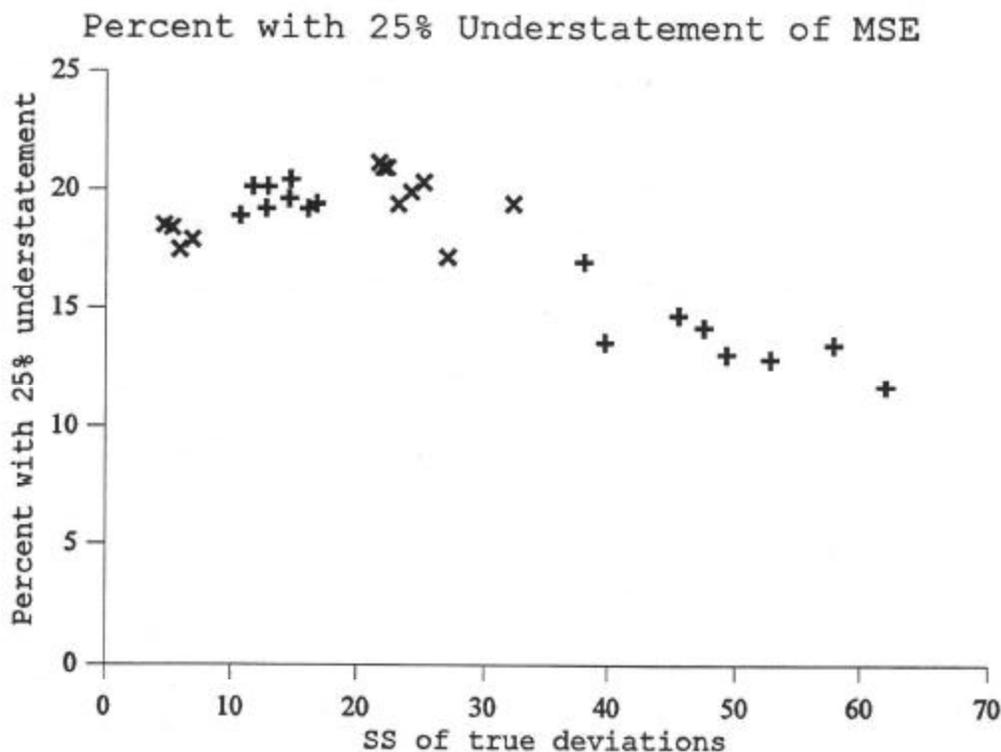


Figure 6 Percent bias in Prasad-Rao MSE estimator, estimated variances, MLE,  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 7 shows the effect on 25% understatement of the actual MSE when the sampling errors are estimated. Compared to Figure 5, the results are much flatter, in the range of 15-20%, compared to the much more dramatic swings in Figure 5. Unlike Figures 3 and 5, the combination of the extra variability from estimating the sampling variances and the somewhat larger actual MSE eliminates the phenomenon of the dramatic drop towards 0% at the right end of the scale.

It was previously noted that REML applied to the sample data and estimated sampling variances yielded estimates with the largest actual MSE. Specifically, choice 5), with the estimator from Cressie (1992), appears here, although it was previously noted that the alternative 6) produces essentially identical results. Figure 8 shows the bias in the estimated MSE for REML. Figure 8 closely resembles Figure 6 in shape but has estimated biases moved up by

roughly 5-10 percentage points. Again, results of this comparison to MLE are consistent with a greater downward bias in the MSE for the latter.



**Figure 7** Percent understatement of true MSE by 25 percent or more, estimated variances, MLE,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

Figure 9 resembles Figure 7, in showing a flatter performance over the range than Figures 3 and 5. Overall, however, the comparison of Figure 9 to Figure 7 awards a significant advantage to REML compared to MLE in preventing marked understatement of the true MSE. This finding is consistent with the relative shift in bias of the MSE estimators compared in Figures 6 and 8.

As noted earlier, use of sample variances in the method of moments estimator produces an increase in actual MSE comparable to the increase for MLE. Figure 10 shows performance comparable or slightly better than that of REML in Figure 8 under the same

circumstances. Again, the MSE estimates exhibit less downward bias than for MLE in Figure 6.

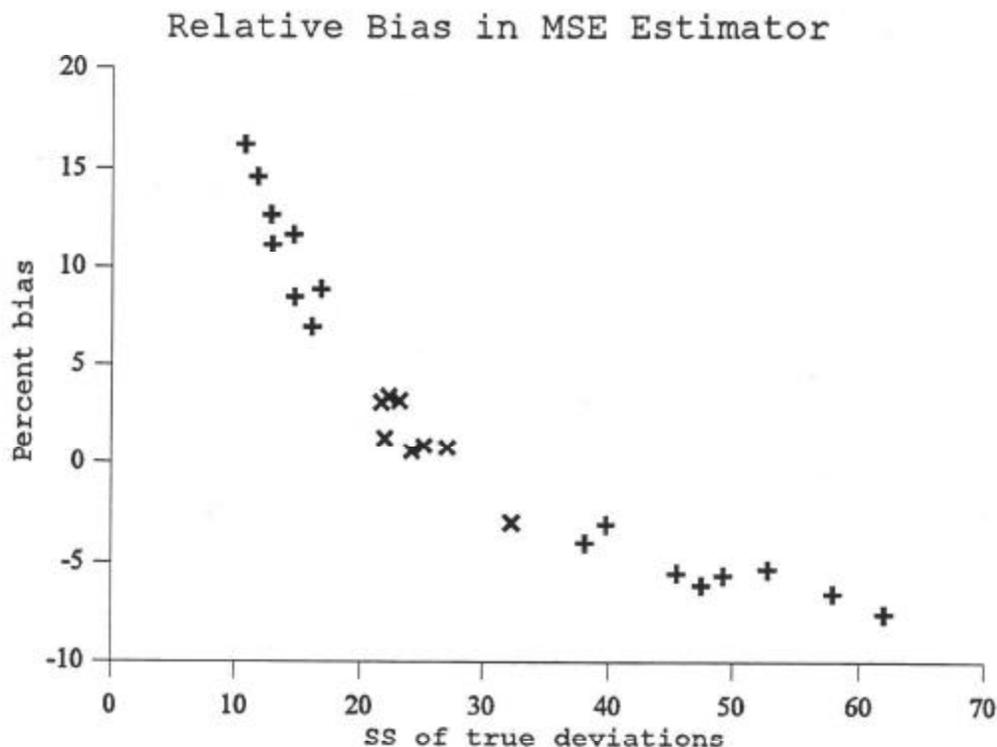
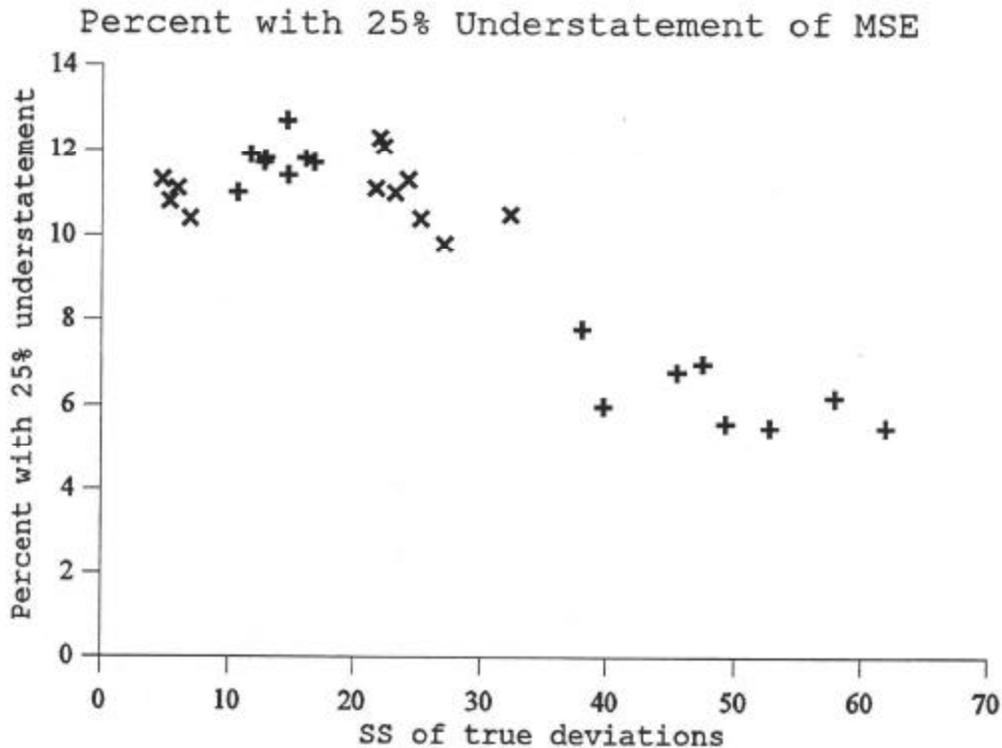


Figure 8 Percent bias in Prasad-Rao MSE estimator, estimated variances, REML,  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

Comparison of Figures 11 and 9 reveals that the slight bias advantage of the method of moments approach compared to REML, shown previously by Figures 10 and 8, is traded against more frequent understatement of the actual MSE by 25% or more. Consequently, there is not a single winner in the contest of these alternatives.

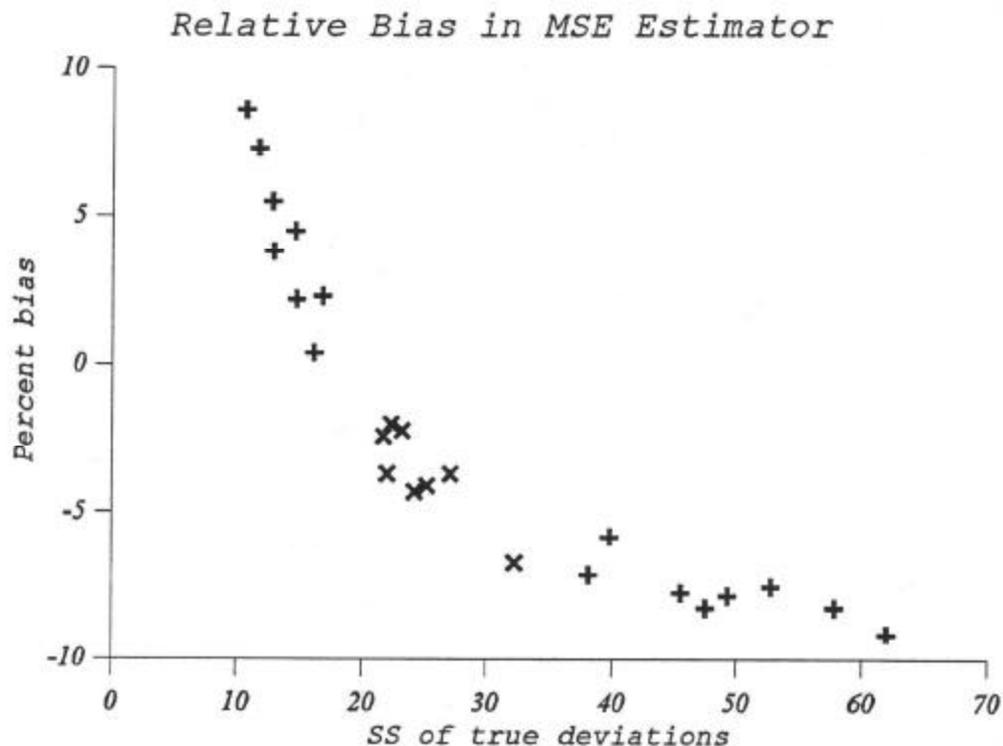
Generally, the method of moments approach does appear to outperform MLE in Figures 6 and 7. The method of moments is subject to less downward bias than MLE at the upper end of the range studied and exhibits less frequent understatement.



**Figure 9** Percent understatement of true MSE by 25 percent or more, estimated variances, REML,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

Figures 12 and 13 present the results for the last alternative, 8), which weights observations equally in estimating  $\sigma^2$  and which does not require iteration. The findings show a considerable downward bias in MSE estimation under these conditions. For example, comparison of Figure 12 to Figure 10 shows a more consistent downward bias over much of the range studied. In turn, the probability of 25% understatement is higher in Figure 13 than Figure 11.

Generally, the findings show that the properties of the MSE estimators are affected to a significant degree as a result of estimating sampling variances when there are relatively few observations or degrees of freedom in each of the domains. These empirical findings do not appear to be a straightforward consequence of the available theoretical results.

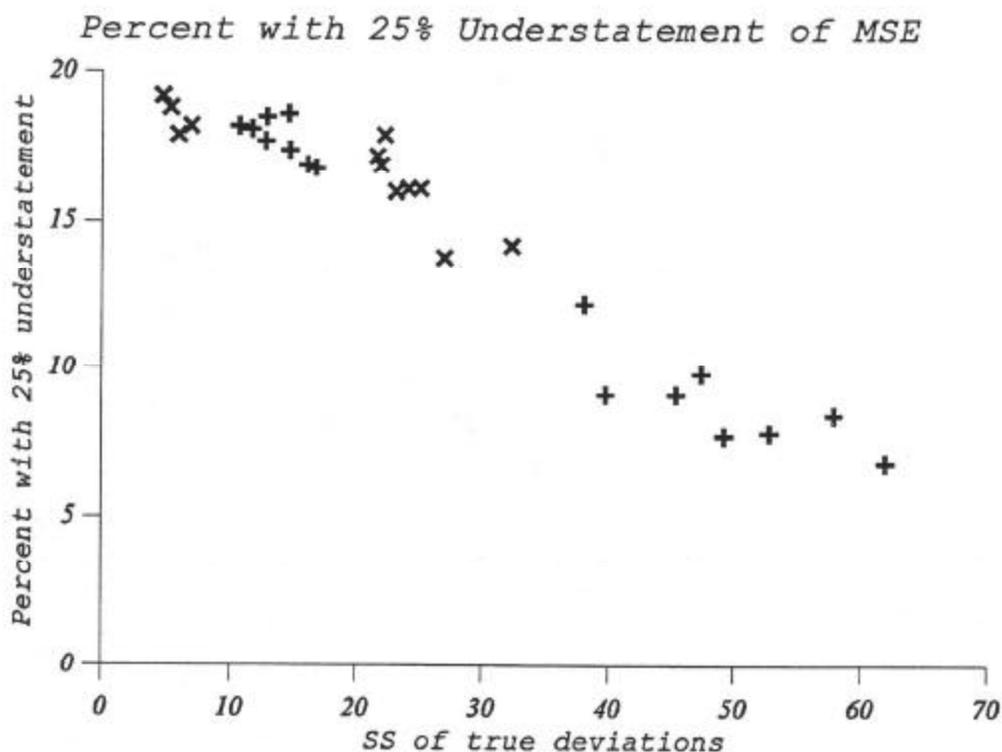


**Figure 10** Percent bias in Prasad-Rao MSE estimator, estimated variances, method of moments,  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

When  $n = 20$  observations are instead available for variance estimation within each cluster, the effects of estimating the variances becomes less pronounced. In other words, the corresponding Figures 6 and 7 for  $m = 20$  become more like Figures 4 and 5, and the pairs of Figures 8 and 9, 10 and 11, and 12 and 13 each resemble Figures 2 and 3 more closely. Consequently, and not surprisingly, the effect on MSE estimation depends on the degree of precision of the sampling variances in the domain, and not simply on the fact that the sampling variances have been estimated.

Translation of the implication of these results to application will, in the author's opinion, not be simple. Compared to the estimation of variance for standard estimates, such as the sample mean, the issue of the variance of the variance, that is, the design-based variance of a variance estimator, is a fairly arcane subject that has consequently received relatively little attention.

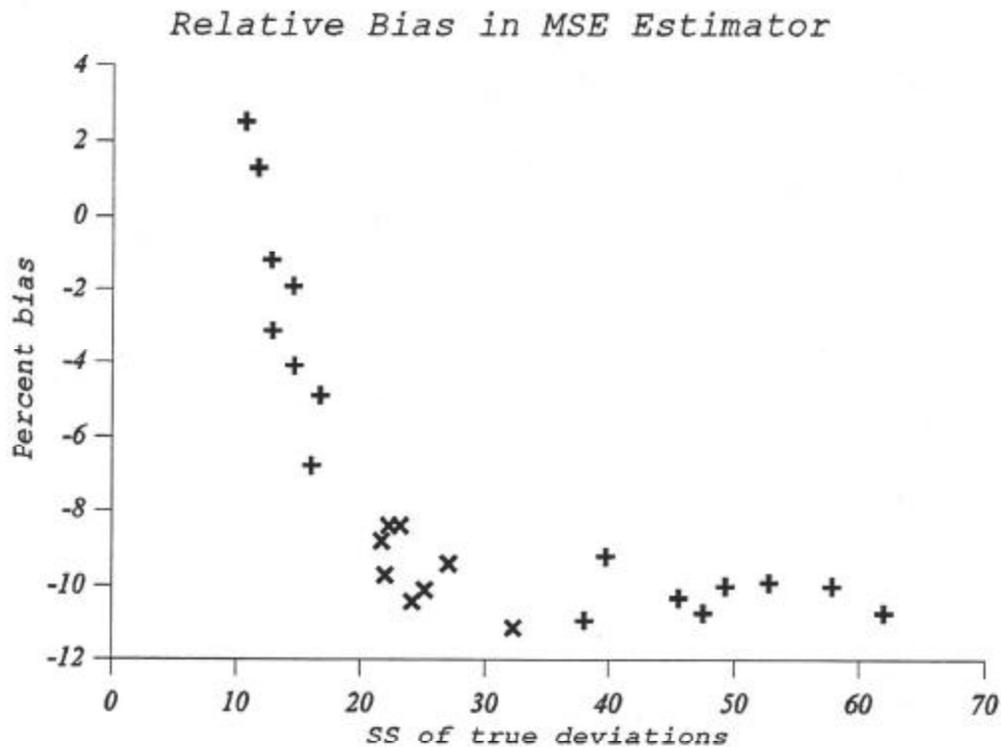
A simple count of the algebraic degrees of freedom will not typically provide an adequate indication of the expected performance of the variance estimator, except in the sense that a variance estimator based on a small number of observations or clusters is certain to be highly variable. Generally, non-normality of the individual or clustered observations may increase the variance of the variance substantially compared to its behavior under normality.



**Figure 11** Percent understatement of true MSE by 25 percent or more, estimated variances, method of moments,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

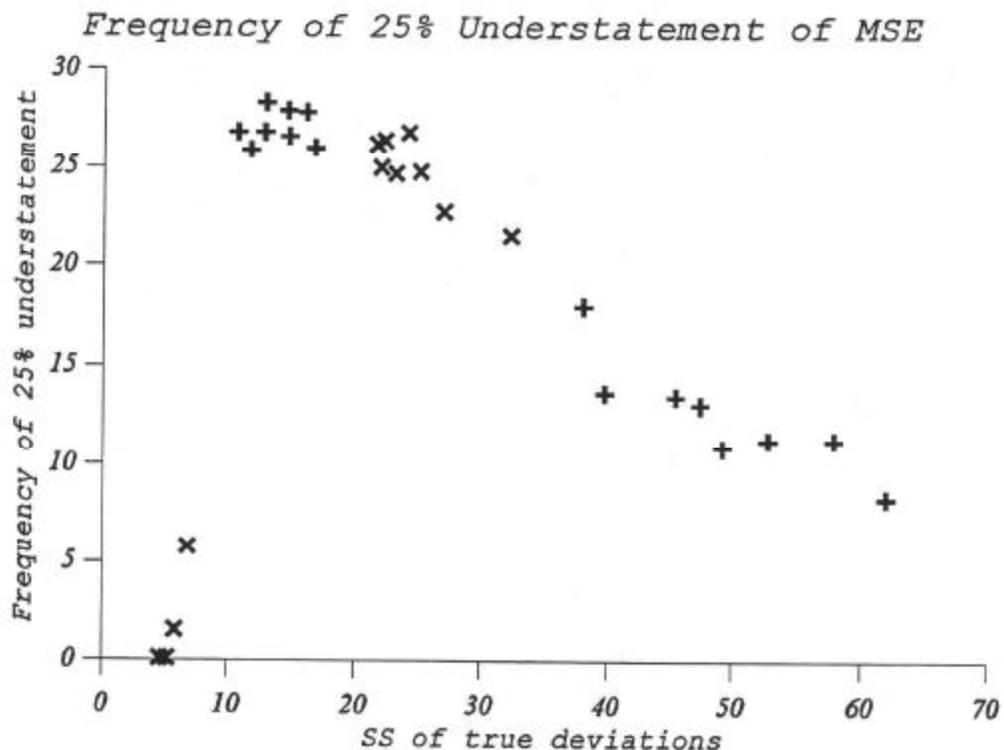
Results for  $m = 20$  domains follow many of the same patterns as  $m = 50$ . Overall, however, there is substantially less evidence to evaluate whether the EBLUP has yielded substantial declines in MSE. When  $m = 50$ , the MSE estimators begin to exhibit relatively extreme behavior, including their upward bias, when the actual reductions are 75% or more. For  $m = 20$ , the same patterns appear much earlier, at around 50% actual reduction. Similarly, the phenomenon

in Figure 3 and others where the MSE estimator suddenly stops overestimating the true MSE by 25% or more shows up much earlier for  $m = 20$ . Thus, effective MSE estimation in situations where the gains from EBLUP are substantial requires numbers of domains on the order of  $m = 50$ . Specific findings are available from the author.



**Figure 12** Percent bias in Prasad-Rao MSE estimator, estimated variances, equally weighted method,  $m = 50$ ,  $p = 4$ , 10 obs. per domain. Note: The first 4 points have been omitted.

Except for separate FORTRAN programs to generate the sample data used in the Monte Carlo study, the variance program VPLX calculated the EBLUP estimators and summarized the results. PC's with 486-class processors performed the calculations for  $m = 20$ , and a Sun SPARC 10 for  $m = 50$ , although selective problems were checked against each other to verify independence of results on the choice of platform.



**Figure 13** Percent understatement of true MSE by 25 percent or more, estimated variances, equally weighted method,  $m = 50$ ,  $p = 4$ , 10 obs. per domain.

### 5. Concluding Remarks

Continued advances in computer technology is certain to have a continued impact on the practice of statistics. Figures 1-13 summarize empirical results that the author would not have had the resources to undertake even a few years ago. Even so, such answers are not yet easily obtained -- for example, each set of points appearing in Figures 1-13 represents about 5 1/2 hrs. of calculation.

The findings, although not generally remarkable, illustrate the subtleties of applying complex estimation methods to practical problems. Features appear that are difficult to anticipate from knowledge of the theoretical results alone. Over time, Monte Carlo

assessment should become even more of a standard to complement theoretical findings.

Substantially more work can and should be done. Section 4.1 outlines a general strategy for useful additional study. As examples, the effect of linkage between  $\theta_i$  and  $\psi_i$  can and should be studied in this manner. Variance generalization has appeared in applications, but what are the consequences of applying a deficient model, i.e., a variance generalization that overpredicts some sampling variances and underpredicts others? What are the consequences of misspecifying (2.1)? How should the variance effects of missing data be taken into account? Issues such as these may have a substantial effect on the behavior of EBLUP procedures, and further Monte Carlo work offers an effective approach.

#### REFERENCES

- Cressie, N. (1992), "REML Estimation in Empirical Bayes Smoothing of Census Undercount," *Survey Methodology*, 18, 75-94.
- Ericksen, E.P., and Kadane, J.B. (1985), "Estimating the Population in a Census Year (with discussion)," *Journal of the American Statistical Association*, 80, 98-131.
- Fay, R.E. (1992), "Inference for Small Domain Estimates from the 1990 Post Enumeration Survey," unpublished manuscript.
- 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.
- 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," W. L. Schaible and M. A. Gonzalez (eds.), *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget, pp. 9-1 - 9-17.
- Ghosh, M., and Rao, J.N.K. (1994), "Small Area Estimation: An Appraisal (with discussion)," *Statistical Science*, 9, 55-93.
- Henderson, C.R. (1950), "Estimation of Genetic Parameters (abstract)," *Annals of Mathematical Statistics*, 21, 309-310.
- 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*, 78, 47-59.

Patterson, H.D., and Thompson, R. (1971), "Recovery of Interblock Information When Block Sizes are Unequal," *Biometrika*, 58, 545-554.

\_\_\_\_\_ (1974), "Maximum Likelihood Estimation of Components of Variance," *Proceedings of the 8th International Biometric Conference*, Washington, D.C., Biometric Society, pp. 197-207.

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.

Schaible, W.L. and Gonzalez, M.A., eds. (1993), "Indirect Estimators in Federal Programs," *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget.

## DISCUSSION

Phillip S. Kott  
National Agricultural Statistics Service

For a number of years now, many of us in the survey sampling community have been grappling with the following question:

"What is the proper role of models in survey sampling?"

The answer for survey sampling purists can be found in Hansen, Madow, and Tepping (1983). Their Guiding Principle No. 4 states:

"Models are appropriately used to guide and evaluate the design of probability samples [including the choice of estimators], but with large samples the inference should not depend on the model."

This principle clearly justifies the use of model-assisted methods within a randomization-based framework, which is the basis for Särndal, Swensson, and Wretman's celebrated new textbook (1992). It is in sharp contrast, however, to the approach that Hansen and his colleagues label "model-dependent."

Unfortunately, it is not at all clear how Guiding Principle No. 4 applies to the issue of estimation in small domains. In fact, in Guiding Principle No. 7, Hansen, Madow, and Tepping concede:

"... model-dependent methods may have an advantage with quite small samples, for which probability-sampling may not be appropriate"

This suggests that our original question needs to be turned around:

"What is the proper role of randomization-based inference when estimating small domains?"

To Bayesians like Don Malec and Joe Sedransk, the answer to this question is simple: "none." Others, like Bob Fay and myself, would like to estimate the value  $\theta_i$  in a small domain  $i$  with an estimator  $t_i$  that has the following property: as the sample size within domain  $i$  grows arbitrarily large (but the sampling fraction stays fixed),  $t_i$  approaches  $\theta_i$  in probability irrespective of the validity of the model used in choosing  $t_i$ .

We realize, of course, that the sample size within domain  $i$  is not arbitrarily large. In fact, in small domain estimation, the sample size within  $i$  is usual so small that a conventional model-

assisted, randomization-based estimator,  $t_{i(rb)}$ , has an unacceptably large standard error, hence the need for a more creative, small domain estimator in the first place! Still, we would not be happy using an estimator that did not work well when it should; that is, when the sample size within domain  $i$  was large.

One can write

$$t_{i(rb)} = \theta_i + s_i, \quad (1)$$

where  $s_i$  is the sampling error of estimator  $t_{i(rb)}$ . Let us assume that the model-assisted randomization-based estimator  $t_{i(rb)}$  is (at least) nearly randomization unbiased so that  $E_p(s_i) \approx 0$ , where the subscript  $p$  denotes that the expectation is with respect to the probability sampling process ("nearly unbiased" means that the bias is small because the sample size across all domains is large). Let us also assume that  $\theta_{i(rb)}$  is nearly unbiased under a model governing the elements of the population; that is to say,  $E_M(s_i) \approx 0$ , where the subscript  $M$  denotes that the expectation is with respect to the model. Finally, let us assume that the  $t_{i(rb)}$  is randomization consistent; i.e.  $\text{plim}_{n(i) \rightarrow \infty} (t_i/\theta_i) = 1$ , where  $n(i)$  is the sample size in domain  $i$ .

In small domain estimation, it is common to model the behavior of the domain values  $\theta_i$  as well as of the population's elements. For convenience, let us restrict our attention to the following domain-level "random effects" model:

$$\theta_i = \mu(\mathbf{x}_i) + \epsilon_i, \quad (2)$$

where  $\mathbf{x}_i$  is a vector of characteristics for domain  $i$ ,  $\mu$  has a known functional form (e.g., linear or logistic) but unknown parameters, and  $\epsilon_i$ , the random effect, is a random variable with mean zero and positive variance.

Let  $m_i$  be a nearly unbiased estimator for  $\mu(\mathbf{x}_i)$ . An estimator for  $t_i$  of the form:

$$t_i^{(g)} = (1 - g)t_{i(rb)} + gm_i$$

is nearly model unbiased. Its mean squared error is (approximately) minimized when

$$g = \frac{\text{Var}(s_i)}{\text{Var}(s_i) + E[(\theta_i - m_i)^2]}. \quad (3)$$

Whether  $\text{Var}(s_i)$  is the model or randomization variance of  $s_i$  depends on whether one's goal is to minimize the model or design variance of  $t_i^{(g)}$ . The same holds true for the interpretation of  $E[(t_i - m_i)^2]$ .

From both a model and randomization-based perspective,  $\text{Var}(s_i)$  and  $E[(\theta_i - m_i)^2]$  are unknown. Särndal, Swensson, and Wretman's estimator for  $\text{Var}_p(t_{i(\text{rb})}) \approx \text{Var}_p(s_i)$  is also a reasonable estimator for the model variance of  $s_i$ . A reasonable estimator for  $E_p[(\theta_i - m_i)^2]$  is illusive, but a good estimator for  $E_\mu[(\theta_i - m_i)^2] \approx \text{Var}(\epsilon_i)$  is not difficult to develop.

Suppose one estimates  $\text{Var}_\mu(s_i)$  and  $\text{Var}(\epsilon_i)$  from the sample, plugs those estimates into equation (3), and then computes  $t_i^{(g)}$ . Call the result  $t_i^*$ . As the sample size in domain  $i$  increases,  $\text{Var}_\mu(s_i)$  decreases, while  $\text{Var}(\epsilon_i)$  remains a positive constant. Thus, as  $n(i)$  grows arbitrarily large  $t_i^*$  converges to  $t_{i(\text{rb})}$ , making it randomization consistent just like  $t_{i(\text{rb})}$ . In fact,  $t_i^*$  is fully in the spirit with Hansen, Madow, and Tepping's Guiding Principle No. 4: models have been used in the choice of the estimator, but the estimator itself, while biased, is randomization consistent.

Let us now turn to the primary question addressed in the Malec & Sedransk and Fay papers: how should the variance of a small domain estimator like  $t_i^*$  be estimated? Both papers take a model-dependent approach. The problem with this approach, of course, is that models can fail. Since Fay's paper deals with simulations, he avoids the problem. Malec & Sedransk do not.

Malec & Sedransk are to be commended for their thoughtful and thorough work in developing complex models at both the element and domain levels that are appropriate for the survey data they are examining. I have absolutely no problems with the determined parts of these models. What bothers me are the random parts. In particular, the authors build in random effects at the county level only. They allow no additional clustering effects within area segments or households. Moreover, they assume county effects are uncorrelated both across adjacent counties and within states. An example of counties in a state likely to be correlated are Kings, Queens, New York, and Bronx Counties -- the four big boroughs of New York City. I suspect that more than one of these counties are represented in the authors' sample.

It should be noted that the goal of the Malec & Sedransk paper is to produce state not county estimators. Their domain-level model is on the county level, however. Thus, they estimate  $\theta_{(\text{state})} = \sum_{i \in \text{state}} \theta_i$  with  $\sum_{i \in \text{state}} t_{i(\text{MS})}$ , where  $t_{i(\text{MS})} = m_i$  for counties not represented in the sample. For counties represented in the sample,  $t_{i(\text{MS})}$  is similar to the  $t_i^*$  discussed above. Nevertheless, because of how the other counties are handled, there is no easy way of modifying a Malec/Sedransk state estimator to make it randomization consistent.

If  $g$  were determined from an outside source, the model variance of  $t_i^{(g)}$  would be

$$\text{Var}_M(t_i^{(g)}) \approx (1 - g)^2 \text{Var}_M(s_i) + g^2 \text{Var}(\epsilon_i). \quad (4)$$

Once estimators for  $\text{Var}_M(s_i)$  and  $\text{Var}(\epsilon_i)$  are computed, an estimator for  $\text{Var}_M(t_i^{(g)})$  quickly presents itself.

When a  $g$  (approximately) satisfying equation (3) is determined from the sample so that  $t_i^{(g)} = t_i^*$ , it is tempting to simply plug that value into equation (4) along with estimates of  $\text{Var}_M(s_i)$  and  $\text{Var}(\epsilon_i)$ . A good deal of high powered statistical work has gone into showing why such a practice can be mistaken. I have a more prosaic problem with this approach to variance estimation: it relies entirely on the truth of the model; in particular, on the model for the  $\epsilon_i$ . It is true that we modeled the  $\epsilon_i$  in developing the estimator  $t_i^*$  in the first place, but to my mind this fact only reinforces a need to be able to evaluate the accuracy of  $t_i^*$  in a way that does not require the same model assumptions.

The randomization mean squared error of  $t_i^{(g)}$  is

$$\text{MSE}_p(t_i^{(g)}) \approx (1 - g)^2 \text{Var}_p(s_i) + g^2 E_p[(\theta_i - m_i)^2].$$

Let  $v(s_i)$  be a randomization-based estimator for  $\text{Var}_p(s_i)$ . One can estimate  $E_p[(\theta_i - m_i)^2]$  with  $(t_{i(\text{rb})} - m_i)^2 - v(s_i)$ . Unfortunately, this estimator is dreadfully unstable. It has, at most, 1 degree of freedom. For many domains,  $v(s_i)$  will also be very unstable, since it has, at most,  $n(i) - 1$  degrees of freedom.

It may come as a shock, but few users of our statistics are all that concerned with variances. With this in mind, perhaps we should abandon the search for a near perfect variance estimator for  $t_i^*$ . We do need to be assured that  $t_i^*$  has some minimum degree of accuracy. One possibility is to model  $v_p(s_i)$  and  $(t_{i(\text{rb})} - m_i)^2 - v(s_i)$  across all the domains and to use the results to derive a conservative indication about the accuracy of  $t_i^*$  for each  $i$ .

#### REFERENCES

Hansen, M., Madow, W. and Tepping, B. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys (with discussion)," Journal of the American Statistical Association, 776-807.

Särndal, C. Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling. New York: Springer-Verlag.

**DISCUSSION OF SMALL AREA ESTIMATION PAPERS  
COPAFS CONFERENCE, MAY 26, 1994**

**David A. Marker  
Westat, Inc.**

Both of these papers are important for their general approach to the problem of small area estimation: they attempt to understand the application of new methods through the explicit use of models. Ideally, one would always design surveys to allow for the production of accurate, direct, design-based estimates. However, when such estimates cannot be produced, one is left with only two choices: either don't produce estimates or use models.

Malec and Sedransk present the use of hierarchical Bayes procedures for small area estimation. I find this approach to be more satisfactory than empirical Bayes procedures for at least three reasons. First, hierarchical Bayes procedures do not assume a particular model to be true. To quote George Box, "All models are wrong, but some are useful." Second, by assuming instead that the truth comes from within a class of prior distributions, it is possible to examine the robustness of the estimates; although this is limited to the range of priors contained in the class. Third, hierarchical Bayes allows for the use of informative priors. While Malec and Sedransk do not make use of informative priors, this is a possible area for extending their results. Particularly for repeated surveys such as the NHIS, there is a wealth of historical data that can be used. These data can be incorporated for model selection, as variables in the actual model, or to construct informative priors.

Many authors, including Malec and Sedransk, use the Gibbs sampler to produce hierarchical Bayes estimates. The advantage of the Gibbs sampler is that it allows for computations from complex distributions. However, the experiences relayed by Malec and Sedransk and others indicate that this approach is extremely time intensive, in some cases taking months to produce stable estimates. This raises questions about the actual utility of this approach to produce timely small area estimates.

One additional point is worth making regarding the Gibbs sampler. As mentioned earlier there is an abundance of historical NHIS data from which informative priors could be developed. It would be very interesting to see the results of using the Gibbs sampler when beginning with informative, rather than uninformative, priors.

Malec and Sedransk develop their model using forward stepwise regression. While this is a reasonable approach, it can lead to suboptimal results under complex situations. Therefore, it might be worthwhile to examine alternative model-selection methods. In selecting their model, they disregarded the sampling weights. They reported that the weights would not have had significant impact based on analyses at the national level. My concern is that given state-to-state differences this might not imply that nothing is lost by disregarding weights when producing state estimates.

Fay uses simulation to examine the real situation of computing the accuracy of small area estimates when the variances are unknown. The Prasad and Rao approach that he evaluates is limited to situations in which the mean and variance are independent. Unfortunately in many situations, including the binomial variable used by Malec and Sedransk, this is not true. Prasad and Rao developed a procedure for producing approximately unbiased mean square errors (MSEs) for model-dependent small area estimates. These MSE estimates are, however, conditional on the model.

For government agencies there is a strong interest in producing design-based measures of accuracy, not ones conditional on models. A method for producing design-based small area specific MSEs was introduced by Marker (1993). This approach replaces the average MSE of Gonzalez and Waksberg with a small area specific MSE, where the variance of the model-dependent estimator is computed for each small area  $i$  using replicated methods (jackknife or balanced repeated replication). The bias is computed by averaging across small areas.

$$MSE(y_i) = var(y_i) + avebias^2(y_i)$$

where

$$avebias^2(y_i) = aveMSE(y_i) - avevar(y_i)$$

This estimator is not completely small area specific, but if the variance term dominates the bias, the root mean square error will provide a useful substitute for the traditional standard error. If the bias term dominates, the small areas can be grouped by expected similar biases. The average bias can then be computed separately for each group

of areas so that the MSE more accurately reflects small area differences. It would be very useful if both Fay and Malec and Sedransk could examine the utility of this approach.

Reference:

Marker, David A., "*Small Area Estimation for the National Health Interview Survey*," Proceedings of the American Statistical Association Section on Survey Research Methods, 1993.