# Session 5

## Small Area and Longitudinal Estimation Using Information from Multiple Surveys

# Small Area and Longitudinal Estimation Using Information from Multiple Surveys

**Sharon L. Lohr**
Department of Mathematics and Statistics, Arizona State University

## 1. Introduction

Most large sample surveys conducted by agencies such as the U.S. Bureau of the Census provide accurate statistics at the national level. Many policymakers and researchers, however, also want to obtain statistics for smaller domains such as states, counties, school districts, or demographic subgroups of a population. These domains are called small areas—so called because the sample size in the area or domain from the survey is small. The goal is to estimate $\theta_i$, the mean value (or other characteristic) of a variable of interest $y$ in small area $i$, for some or all of the small areas.

Small area estimates of income and poverty are employed in the allocation of more than eight billion dollars each year in the U.S. In that setting, no single source of information currently being collected is capable of producing reliable estimates of the number of poor people under age 18 in each county, or the number of poor children in each school district. Thus, the current practice to estimate poverty at the state level (see National Research Council, 2000, p. 49) uses auxiliary information from tax returns, food stamp programs, and the decennial census to supplement the data from the Current Population Survey (CPS). The model used is based on that in the pioneering paper by Fay and Herriot (1979). Let $\theta_i$ be the proportion of school-age children who are poor in state $i$. The direct estimate $\bar{y}_i$ of $\theta_i$ is calculated using data exclusively from the CPS, and $\hat{V}(\bar{y}_i)$ is an estimate of the variance of $\bar{y}_i$. A regression model for predicting $\theta_i$ using auxiliary information is

$$\theta_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji} + v_i \qquad (1)$$

where the $x_{ji}$'s represent covariates for state $i$ (e.g., $x_{2i}$ is the proportion of people receiving food stamps in state $i$) and $v_i$ (assumed to follow a $N(0, \sigma_v^2)$ distribution) is the model error for state $i$. The regression parameters and $\sigma_v^2$ may be estimated using maximum likelihood. The predicted value from the regression equation for state $i$ is combined with the direct estimate $\bar{y}_i$ from the CPS according to the relative amounts of information present in each estimate:

$$\hat{\theta}_i = \hat{\gamma}_i \bar{y}_i + (1 - \hat{\gamma}_i)(\hat{\alpha}_0 + \sum_{j=1}^{k} \hat{\alpha}_j x_{ji}), \qquad (2)$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 / [\hat{\sigma}_v^2 + \hat{V}(\bar{y}_i)]$. If the direct estimate is precise for a state, i.e., $\hat{V}(\bar{y}_i)$ is small, then $\hat{\gamma}_i$ is close to one and $\hat{\theta}_i$ relies mostly on the direct estimate. Conversely, if the CPS contains little information about state $i$'s poverty rate, then $\hat{\gamma}_i$ is close to zero and $\hat{\theta}_i$ relies mostly on the predicted value from the regression. The estimator in (2) generally has smaller mean squared error (MSE) than the direct estimator $\bar{y}_i$ because it uses information available from other sources. In the extreme case where area $i$ has no observations from the CPS and hence $\bar{y}_i$ cannot be calculated, the improvement in MSE is infinite.

Traditionally, as is done for state estimates of school-age poverty, small area estimation relies on a model relating the responses of interest in the small areas to each other and to covariates. The model allows the estimate of $\theta_i$ to "borrow strength" from other small areas through random effects terms and regression parameters. Small area estimation models have been used in many settings to obtain more accurate estimates for subpopulations without additional cost for data collection. A thorough review of research in small area estimation is given in Rao (2003).

As detailed in Rao (2003), two main types of models are used in small area estimation, distinguished by the nature of the auxiliary information. The model described above for estimating poverty rates is an example of an *area-level model*: $\bar{y}_i$, the estimate of $\theta_i$ from the survey, is related to area-level covariates. In an area-level model, the auxiliary information does not need to be known for individual persons in area $i$, since the covariates are summary information for the small areas. In a *unit-level model*, the response of interest for each person in area $i$ is modeled as a function of covariates available for that person. A unit-level model might, for example, model log(income for $j^{th}$ person in area $i$) using covariates of tax return and food stamp data for that person. The unit-level model thus requires that the covariate values are known (and can be linked to the income data) for the persons in the survey.

Both unit- and area-level models assume that the model covariates are measured without error. In many situations, though, auxiliary information is available that can help in the estimation, but that information is not exact. Auxiliary information may be available from another survey, or from an administrative source in which imputation has been used to fill in missing values. In both of these cases, the auxiliary information is measured with error—sampling and nonsampling error for survey data, and imputation error for incomplete administrative data. For example, the American Community Survey (ACS) will sample about 3 million households each year. For most small areas, the ACS will give relatively precise estimates of quantities it measures, and thus can be used as auxiliary information for estimating small area characteristics on many topics. The ACS still contains sampling error for many small areas, however, and that error should be included in standard errors reported for the estimates.

For another example, the U.S. National Crime Victimization Survey (NCVS) provides reliable estimates of victimization rates for the country as a whole. If separate estimates of victimization rates are desired for each state, however, some states have very small sample sizes, and standard errors using a direct estimate are unacceptably large. The same problem occurs when one desires to estimate characteristics of subgroups of the population such as victims of domestic violence—the sample sizes of domestic violence victims are not sufficiently large to give adequate precision for estimates of interest (Ybarra and Lohr, 2002). The Uniform Crime Reports (UCR), which provides statistics compiled by the FBI from law enforcement agencies, could be used as auxiliary information; Wiersema et al. (2000) found high correlations between NCVS and UCR estimates of number of victimizations using data from ten metropolitan statistical areas (MSAs). The UCR data, however, have many limitations. They only include crimes known to police; moreover, reporting is voluntary so many agencies have missing data. Even when agencies do report the data, reporting is not uniform. Maltz (1999) discussed the extent of missing data in the UCR, and described some current imputation schemes. For the UCR to be used as auxiliary information to the NCVS, imputation errors need to be incorporated into estimates of precision.

Many survey designs in the U.S. are now being integrated to allow combination of estimates. The U.S. National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) currently share the same primary sampling units (psu's): the psu's selected for NHIS are used as a sampling frame for NHANES. NHIS is a stratified multistage probability sample of about 100,000 persons (40,000 households) per year. The design is described in detail in Botman et al. (2000). NHANES conducts medical examinations of participants, however, and the mobile examination unit can only visit 15 psu's per year (about 5000 persons), as opposed to 358 psu's for NHIS. Because of the small sample size, NHANES data are usually accumulated over time in order to produce estimates. The small sample sizes also cause state and local estimates from NHANES to have low precision. The NHIS data provide more precise estimates of quantities measured at some localities, but the data come from an interview rather than an examination: For example, in NHANES, prevalence of diabetes may be estimated using the results of the medical exams, while in NHIS respondents are asked questions about health problems. We would expect, though, that the questionnaire results would be highly correlated with the medical examination results, and thus that the NHIS would provide high-quality auxiliary information for use with NHANES data for improved small area estimation.

The following situation is considered in this paper. Suppose there are $t$ areas of interest (for example, $t = 50$ if states are small areas). We are interested in a characteristic $\theta_i$ of area $i$, for $i = 1, \ldots, t$. We have data from the primary survey for some (or all) areas, and data from an auxiliary survey for some (or all) areas. Often the characteristic of interest will be a mean or proportion. For estimating state victimization rates, $\theta_i$ might be the proportion of persons who are victims of violent crime in state $i$. The NCVS is considered the primary survey, and the UCR can be used to provide auxiliary information (although with error). The main questions to be considered for incorporating auxiliary information with error into small area estimates are: (1) How should the information be used in a small area model? and (2) How does the error in the auxiliary information affect the MSE of the small area estimates?

In this paper, we summarize some of our recent research on combining information from surveys to obtain more accurate estimates at the small area and national level. In Section 2, we discuss unit-level models for combining information, and in Section 3 we discuss area-level models that allow for uncertainty in the auxiliary information. Section 4 presents recent work on estimation in multiple frame surveys that can be used in small area estimation, and Section 5 discusses directions for future work.

## 2. Unit-level Models for Use with Multiple Surveys

Lohr and Prasad (2003) developed a framework for combining information from multiple surveys when information is available at the unit level. Let $y_{ij}$ denote the characteristic of interest for the $j^{th}$ unit in area $i$. Let $\mathbf{x}_{ij} = (x_{ij1} \cdots x_{ijk})^T$ denote a vector of other characteristics for unit $j$ of area $i$. For estimating assault rates, $y_{ij}$ might be the number of assaults that would be reported to the NCVS by person $j$ in small area $i$ over a specified time period, and $x_{ij1}$ the number of assaults for that person that would be included in the UCR for the same time period. For estimating income, $y_{ij}$ might be the log of income of household $j$ in area $i$ (measured in the CPS), and $\mathbf{x}_{ij}$ might be related quantities asked in the ACS. In addition, there may exist various covariates $a_{ijl}$ that come from administrative records.

The above paragraph described applications in which $y$ and $\mathbf{x}$ are measured from different surveys. However, the methods also apply to the "sampling on two occasions" setting. Many surveys such as the NCVS have a panel design in which the same households are sampled during several administrations of the survey. In this setting, $y$ may be taken as the value of a characteristic on the second occasion and the auxiliary variable $\mathbf{x}$ is the same variable for the first occasion.

In area $i$, both $\mathbf{x}$ and $y$ are measured on the $n_i^{xy}$ units in $\mathcal{S}_{ixy}$; $\mathbf{x}$ (but not $y$) is measured on the $n_i^x$ units in the set $\mathcal{S}_{ix}$; $y$ (but not $\mathbf{x}$) is measured on the $n_i^y$ units in the set $\mathcal{S}_{iy}$. If unit $(ij)$ in the population is included in both surveys, $m = k + 1$ measurements are recorded.

We use a multivariate mixed model to describe the relationship between $\mathbf{x}$, $y$, and covariates. We assume that observations in different small areas are independent. To simplify expression of results, we assume that the multivariate response vector $\mathbf{u}_i$ is arranged with all observations from $\mathcal{S}_{ixy}$ first, followed by those from $\mathcal{S}_{ix}$ and $\mathcal{S}_{iy}$, so

$$\mathbf{u}_i^T = [\mathbf{x}_{i1}^T, y_{i1}, \ldots, \mathbf{x}_{i,n_i^{xy}}^T, y_{i,n_i^{xy}}, \mathbf{x}_{i,n_i^{xy}+1}^T, \ldots, \mathbf{x}_{i,n_i^{xy}+n_i^x}^T, y_{i,n_i^{xy}+n_i^x+1}, \ldots, y_{i,n_i^{xy}+n_i^x+n_i^y}].$$

Let

$$\mathbf{u}_i = \mathbf{A}_i\boldsymbol{\mu} + \mathbf{Z}_i\mathbf{v}_i + \mathbf{e}_i \tag{3}$$

where $\boldsymbol{\mu}$ is a vector of fixed effects parameters, $\mathbf{A}_i$ and $\mathbf{Z}_i$ are known matrices, and $\mathbf{v}_i$ and $\mathbf{e}_i$ are independent random vectors with mean $\mathbf{0}$. $\mathrm{Cov}\,(\mathbf{v}_i) = \boldsymbol{\Sigma}_v$ and

$$\mathrm{Cov}\,(\mathbf{e}_i) = \mathbf{R}_i = [\mathbf{I}_{n_i^{xy}} \bigotimes \boldsymbol{\Sigma}_e] \bigoplus [\mathbf{I}_{n_i^x} \bigotimes \boldsymbol{\Sigma}_{exx}] \bigoplus [\mathbf{I}_{n_i^y} \bigotimes \boldsymbol{\Sigma}_{eyy}],$$

where the matrices $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e$ are partitioned as

$$\boldsymbol{\Sigma}_v = \begin{bmatrix} \boldsymbol{\Sigma}_{vxx} & \boldsymbol{\Sigma}_{vxy} \\ \boldsymbol{\Sigma}_{vxy}^T & \Sigma_{vyy} \end{bmatrix}, \quad \boldsymbol{\Sigma}_e = \begin{bmatrix} \boldsymbol{\Sigma}_{exx} & \boldsymbol{\Sigma}_{exy} \\ \boldsymbol{\Sigma}_{exy}^T & \boldsymbol{\Sigma}_{eyy} \end{bmatrix}$$

and where $\bigoplus$ represents direct sum and $\bigotimes$ represents Kronecker product. Thus

$$\mathbf{V}_i = \mathrm{Cov}\,(\mathbf{u}_i) = \mathbf{R}_i + \mathbf{Z}_i\boldsymbol{\Sigma}_v\mathbf{Z}_i^T \tag{4}$$

with

$$\mathbf{Z}_i = \left[ \begin{array}{lcl} \mathbf{1}_{n_i^{xy}} & \otimes & \mathbf{I}_m \\ \mathbf{1}_{n_i^x} & \otimes & (\mathbf{I}_k \ \mathbf{0}_k) \\ \mathbf{1}_{n_i^y} & \otimes & (\mathbf{0}_k^T \ 1) \end{array} \right]$$

where $\mathbf{1}_j$ is a $j$-vector of ones.

For simplicity of presentation, we take $\boldsymbol{\mu}$ to be the $m$-vector of fixed effects means, partitioned as $\boldsymbol{\mu}^T = [\boldsymbol{\mu}_x^T \ \mu_y]$. However, all results are easily extended to the case where $\boldsymbol{\mu}$ is a general vector of parameters, and $\mathbf{A}_i$ is a matrix of fixed effects covariates. In this way information from a census or from administrative records may be incorporated into the small area estimates through regression.

Under this setup, Lohr and Prasad (2003) showed that if $\boldsymbol{\mu}$ and the covariance component matrices are known, then the best linear unbiased predictor (BLUP) for $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{ix}^T, \theta_i)^T$ is

$$\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\mu} + n_i^{xy}\mathbf{M}_i\boldsymbol{\Sigma}_e^{-1}(\bar{\mathbf{u}}_{ixy} - \boldsymbol{\mu}) + \mathbf{M}_i\mathbf{n}_i^*(\boldsymbol{\Sigma}_e^*)^{-1}(\bar{\mathbf{u}}_i^* - \boldsymbol{\mu}). \tag{5}$$

Here, $\bar{\mathbf{u}}_{ixy}$ is the average of the $n_i^{xy}$ vectors $(\mathbf{x}_{ij}^T, y_{ij})^T$ for $j \in \mathcal{S}_{ixy}$; $\bar{\mathbf{u}}_i^* = (\bar{\mathbf{x}}_{ix}^T, \bar{y}_{iy})^T$ contains the averages of the $\mathbf{x}_{ij}$'s for $j \in \mathcal{S}_{ix}$ and of the $y_{ij}$'s for $j \in \mathcal{S}_{iy}$;

$$\boldsymbol{\Sigma}_e^* = \left[ \begin{array}{cc} \boldsymbol{\Sigma}_{exx} & 0 \\ 0 & \boldsymbol{\Sigma}_{eyy} \end{array} \right], \tag{6}$$

$$\mathbf{n}_i^* = \left[ \begin{array}{cc} n_i^x\mathbf{I} & 0 \\ 0 & n_i^y \end{array} \right], \tag{7}$$

and

$$\mathbf{M}_i = (\boldsymbol{\Sigma}_v^{-1} + n_i^{xy}\boldsymbol{\Sigma}_e^{-1} + \mathbf{n}_i^*(\boldsymbol{\Sigma}_e^*)^{-1})^{-1}. \tag{8}$$

This estimator reduces to the multivariate estimator in Datta et al. (1999) if $n_i^x = n_i^y = 0$.

The BLUP $\tilde{\theta}_i$ for $\theta_i$ is the $m^{th}$ component of $\tilde{\boldsymbol{\mu}}_i$, and $\text{MSE}(\tilde{\theta}_i) = \mathbf{M}_{iyy}$, the $(m, m)$ entry of $\mathbf{M}_i$. As a special case, the BLUP of $\theta_i$ when $n_i^{xy} = n_i^y = 0$ is $\tilde{\theta}_i = \mu_y + \boldsymbol{\Sigma}_{vxy}^T\boldsymbol{\Sigma}_{vxx}^{-1}(\tilde{\boldsymbol{\mu}}_{ix} - \boldsymbol{\mu}_x)$: the estimator then borrows strength by using the between-area covariance of $\mathbf{x}$ and $y$.

If the quantities from the two surveys are correlated, $\tilde{\theta}_i$ is more efficient than the corresponding estimator that does not use the auxiliary survey data. Lohr and Prasad (2003) derived the gain in efficiency, and showed that $\tilde{\theta}_i$ has smaller MSE than the estimator from the univariate unit-level model of Battese et al. (1988) if $n_i^x n_i^{xy}\boldsymbol{\Sigma}_{exy} \neq 0$ or $n_i^x\boldsymbol{\Sigma}_{vxy} \neq 0$.

## 2.1. Estimation of Unknown Quantities

The estimator in (5) was calculated assuming that the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_v$, and $\boldsymbol{\Sigma}_e$ are known. In practice, these must be estimated from the data.

Using the generalized least squares estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, and using consistent estimators of the covariance components, the multivariate estimator becomes

$$\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\mu}} + n_i^{xy}\hat{\mathbf{M}}_i\hat{\boldsymbol{\Sigma}}_e^{-1}(\bar{\mathbf{u}}_{ixy} - \hat{\boldsymbol{\mu}}) + \hat{\mathbf{M}}_i\mathbf{n}_i^*(\hat{\boldsymbol{\Sigma}}_e^*)^{-1}(\bar{\mathbf{u}}_i^* - \hat{\boldsymbol{\mu}}). \tag{9}$$

Lohr and Prasad (2003) derived the second order asymptotic properties of this estimator. As with the BLUP, $\hat{\theta}_i$ is the $m^{th}$ component of $\hat{\boldsymbol{\mu}}_i$.

The method has been implemented in R and S-Plus, with restricted maximum likelihood used to estimate the covariance components. A simulation study demonstrated that $\hat{\theta}_i$ was much more efficient than an estimator that did not use the information from the auxiliary survey, particularly when $\boldsymbol{\Sigma}_{vxy}$ was large relative to $\boldsymbol{\Sigma}_{vxx}$ and $\Sigma_{vyy}$. Even with relatively modest sample sizes in the auxiliary survey (say, $n_i^x = 5$), when the survey quantities were highly correlated the MSE of $\hat{\theta}_i$ was about 1/5 of the MSE of the univariate unit-level estimator that did not use the $\mathbf{x}$ information.

When using the multivariate estimator with separate surveys, in most cases it will not be necessary to match sample observations between the two surveys. Even when the survey designs share the same primary sampling units, it is unlikely that the same persons are included in the surveys. Thus, it is overwhelmingly probable that in most small areas, $n_i^{xy} = 0$. Consequently, the estimator in (5) will involve $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e^*$ but not $\boldsymbol{\Sigma}_{exy}$. The vector $\boldsymbol{\Sigma}_{exy}$ is the only quantity, however, whose estimation requires that units in the two surveys be matched. The matrix $\boldsymbol{\Sigma}_e^*$ can be estimated from the two separate surveys, and $\boldsymbol{\Sigma}_v$ can be estimated provided that the number of small areas that contain observations from both surveys is sufficiently large.

### 2.2. Robust Estimation of Covariance Components

The unit-level multivariate approach depends on a model, and the estimates are therefore sensitive to departures from that model. The estimates of the fixed effects and of the covariance components can perform badly in the presence of aberrant observations. In particular, the restricted maximum likelihood estimates of the covariance components that were used in (9) are affected by outliers. Outliers will not be too great of a problem for estimating $\boldsymbol{\Sigma}_e$ because in most situations there will be sufficient degrees of freedom at the within-area level to mitigate the effect of a few moderate outliers. There are fewer degrees of freedom for estimating $\boldsymbol{\Sigma}_v$, however, so if the estimated mean of a small area is aberrant, this outlying area may greatly affect the REML estimate of $\boldsymbol{\Sigma}_v$.

Dueck and Lohr (2003) developed a method for robust estimation of multivariate covariance components. They used multivariate M-estimation of random effects to reduce the influence of outliers—at both the within-area and between-area levels—on the estimated covariance components. Preliminary research indicates that use of this method, together with robust estimation of the fixed effects, improves the accuracy of small area estimates when some data may be contaminated.

## 3. Area-level Models for Multiple Surveys

The models in Section 2 result in improved efficiency when unit-level auxiliary information exists and observations can be matched across surveys. Matching is easy when sampling on two occasions, where $y$ is the response of interest measured at time 2 and the auxiliary information is the same response measured at time 1. In other settings, only areas may need to be matched, since different units will be used in the two surveys. For some applications, however, matching units may be infeasible: records from the NCVS cannot in general be matched with the same persons' records from the UCR. In addition, there may be concerns that using unit-level data across surveys or other data sources may compromise confidentiality of the data (see Lohr, 2003). For some surveys, respondents may not have given permission to have their data combined with individual-level information from other sources. In such cases, area-level models are preferred.

In this section, we examine area-level models for use with two surveys. To simplify presentation, we consider the case where $\theta_i$ is a population mean, although extensions to other parameters are readily made. Let $\bar{y}_i$ be an unbiased estimator of $\theta_i$ from the primary survey, with sampling variance $V(\bar{y}_i) = \psi_i$. Administrative data for area $i$, $\mathbf{A}_i$, is assumed to be measured without error. We consider the $k$-vector $\mathbf{X}_i$ to be population characteristics for area $i$ which in some areas can be estimated by a vector $\mathbf{x}_i$ from the auxiliary data source. Often, $\mathbf{X}_i$ will be a vector of population means for area $i$. We assume here that when $\mathbf{x}_i$ is measured, $E(\mathbf{x}_i) = \mathbf{X}_i$ and $V(\mathbf{x}_i) = \mathbf{\Sigma}_i$.

### 3.1. What if Error in Auxiliary Information is Ignored?

The Fay-Herriot (1979) model leads to the BLUP of $\theta_i$. If $\bar{y}_i$ and $\theta_i$ are assumed to be normally distributed, the Fay-Herriot estimator can be motivated in an empirical Bayesian framework (see Rao, 2003, chapter 9). It is assumed that $\bar{y}_i \mid \theta_i, \psi_i \sim N(\theta_i, \psi_i)$; a regression model for the population quantity is given by

$$\theta_i|\mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim N(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}, \sigma_v^2). \tag{10}$$

If the quantities $(\bar{y}_i, \theta_i)$ are independent for $i = 1, \ldots, t$, then the posterior distribution of $\theta_i$ is

$$\theta_i|\bar{y}_i, \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \psi_i \sim N[\gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}), \psi_i\gamma_i^*] \tag{11}$$

where $\gamma_i^* = \sigma_v^2/(\sigma_v^2 + \psi_i)$. The mean of the posterior distribution of $\theta_i$ is

$$\tilde{\theta}_{iEB} = \gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{X}_i^T\boldsymbol{\beta}). \tag{12}$$

Now let us examine what happens if an estimator $\hat{\mathbf{X}}_i$ with MSE$(\hat{\mathbf{X}}_i) = \mathbf{C}_i$ is substituted for the population quantity $\mathbf{X}_i$ in (12); either $\mathbf{x}_i$ or another estimator may be used for $\hat{\mathbf{X}}_i$. Let

$$\tilde{\theta}_i^* = \gamma_i^*\bar{y}_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}). \tag{13}$$

Then MSE$(\tilde{\theta}_i^*) = \psi_i\gamma_i^* + (1-\gamma_i^*)^2\boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$, and the posterior variance in (11) underestimates the true variance. If the matrix $\mathbf{C}_i$ is large, the mean squared error of the $\tilde{\theta}_i^*$ can be larger than $\psi_i$, so that the supposedly improved small area estimator can perform worse than the direct estimator that uses no auxiliary information. In addition, if the error in estimating $\mathbf{X}_i$ is ignored and $\psi_i\gamma_i^*$ is naively reported to be the MSE, the estimator will be thought to be more precise than it really is.

We can correct the MSE by incorporating the error in estimating $\mathbf{X}_i$ into the model in (10). If $\mathbf{x}_i|(\mathbf{X}_i, \mathbf{\Sigma}_i) \sim N(\mathbf{X}_i, \mathbf{\Sigma}_i)$, $\mathbf{X}_i|(\boldsymbol{\mu}_x, \mathbf{\Sigma}) \sim N(\boldsymbol{\mu}_x, \mathbf{\Sigma})$, and the quantities $(\mathbf{x}_i, \mathbf{X}_i)$ are independent across areas, then the posterior distribution of $\theta_i$ has mean

$$\gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T\boldsymbol{\alpha} + \mathbf{c}_i^T\boldsymbol{\beta})$$

and variance

$$\psi_i\gamma_i^* + (1-\gamma_i^*)^2\boldsymbol{\beta}^T\mathbf{D}_i\boldsymbol{\beta},$$

where

$$\mathbf{c}_i = \mathbf{\Sigma}(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{x}_i + \mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\boldsymbol{\mu}_x$$

and

$$\mathbf{D}_i = \mathbf{\Sigma}(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{\Sigma}_i = \mathbf{\Sigma}_i(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\mathbf{\Sigma}.$$

With the additional assumptions on the distribution of the auxiliary survey data, the posterior variance is correct for the MSE. The relative weight $\gamma_i^*$, however, still does not account for the error in estimating $\mathbf{X}_i$; it is possible for the posterior variance to be larger than $\psi_i$ so that incorporating the auxiliary $\mathbf{x}$ information may result in a decrease in precision. The methods in the following sections use the uncertainty about $\mathbf{x}_i$ when determining the relative weightings of the direct and indirect estimators.

### 3.2. Multivariate Fay-Herriot Model

Fay (1987) and Datta et al. (1991) developed a Fay-Herriot-type model for a multivariate response, and showed that it often results in more efficient estimators for a small area quantity of interest than the univariate Fay-Herriot model. Datta et al. (1991) were interested in estimating the median income of four-person households in state $i$. The direct estimate was from the CPS. The auxiliary information, $x_i = (3/4)$ (median income of five-person households) + (1/4) (median income of three-person households) also came from the CPS. The multivariate model they used reduced the MSE of the estimator of $\theta_i$ through correlations with the other variables. Lohr and Ybarra (2003) extended this model to allow for missing observations, and to allow the observations to come from different sources. The following summarizes the results for the notationally simpler case when $\mathbf{x}_i$ and $\bar{y}_i$ are independent.

Let $\mathbf{U}_i = [\mathbf{X}_i^T,\ \theta_i]^T$ represent the population values for each of the $i$ areas, $i = 1, \ldots, t$. Define $\mathbf{T}_i$ to be the matrix whose $j^{th}$ row is $[\mathbf{0}^T, \cdots, \mathbf{0}^T, \mathbf{A}_i^T, \mathbf{0}^T, \cdots, \mathbf{0}^T]$ where the $\mathbf{A}_i^T$ occurs as the $j^{th}$ column. Consider the model

$$\mathbf{U}_i = \mathbf{T}_i\boldsymbol{\alpha} + \mathbf{v}_i \tag{14}$$

where $\mathbf{v}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_b)$ and $\boldsymbol{\alpha}$ is a vector of regression coefficients. As in the unit-level model, the covariance matrix $\mathbf{\Sigma}_b$ is partitioned as

$$\mathbf{\Sigma}_b = \begin{bmatrix} \mathbf{\Sigma}_{bxx} & \mathbf{\Sigma}_{bxy} \\ \mathbf{\Sigma}_{bxy}^T & \mathbf{\Sigma}_{byy} \end{bmatrix}.$$

Define the vector $\mathbf{u}_i$ and the matrices $\mathbf{Z}_i$ and $\mathbf{\Psi}_i$ for three cases:

1. If $\mathbf{x}$ and $y$ are both observed for area $i$ then $\mathbf{u}_i = [\mathbf{x}_i^T, \bar{y}_i]^T$, $\mathbf{Z}_i = \mathbf{I}_{k+1}$, and
$$\boldsymbol{\Psi}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0}^T & \psi_i \end{bmatrix}.$$
2. If $\mathbf{x}$ is observed in area $i$ but $y$ is not observed then $\mathbf{u}_i = \mathbf{x}_i$, $\mathbf{Z}_i^T = [\mathbf{I}_k, \mathbf{0}_k]$ and $\boldsymbol{\Psi}_i = \boldsymbol{\Sigma}_i$
3. If $y$ is observed in area $i$ but $\mathbf{x}$ is not observed then $\mathbf{u}_i = \bar{y}_i$, $\mathbf{Z}_i^T = [\mathbf{0}_k^T, 1]$, and $\boldsymbol{\Psi}_i = \psi_i$.

Then the observations $\mathbf{u}_i$ follow the model
$$\mathbf{u}_i = \mathbf{Z}_i^T \mathbf{T}_i \boldsymbol{\alpha} + \mathbf{Z}_i^T \mathbf{v}_i + \mathbf{e}_i, \tag{15}$$
where $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Psi}_i)$. The covariance matrix of $\mathbf{u}_i$ is
$$\mathbf{V}_i = \mathbf{V}_i(\mathbf{u}_i) = \mathbf{Z}_i^T \boldsymbol{\Sigma}_b \mathbf{Z}_i + \boldsymbol{\Psi}_i.$$
The $\mathbf{u}_i$'s are assumed to be independent. This model then fits into the block diagonal covariance structure model described in Section 6.3 of Rao (2003). Define
$$\tilde{\boldsymbol{\alpha}} = \left( \sum_i \mathbf{T}_i^T \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{Z}_i^T \mathbf{T}_i \right)^{-1} \left( \sum_i \mathbf{T}_i^{-1} \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{u}_i \right),$$
$$\mathbf{K}_i = (\boldsymbol{\Sigma}_{bxx} + \boldsymbol{\Sigma}_i)^{-1},$$
and
$$\kappa_i = \frac{\Sigma_{byy} - \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \boldsymbol{\Sigma}_{bxy}}{\Sigma_{byy} - \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \boldsymbol{\Sigma}_{bxy} + \psi_i},$$
The BLUP for $(\mathbf{X}_i^T, \theta_i)$ is then
$$\tilde{\theta}_{iMFH} = \kappa_i \bar{y}_i + (1 - \kappa_i) \left[ [\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i \left( \mathbf{x}_i - [\mathbf{I}, \mathbf{0}] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} \right) \right] \tag{16}$$
if both $\mathbf{x}_i$ and $\bar{y}_i$ are observed in area $i$;
$$\tilde{\theta}_{iMFH} = [\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} + \boldsymbol{\Sigma}_{bxy}^T \mathbf{K}_i (\mathbf{x}_i - [\mathbf{I}, \mathbf{0}] \mathbf{T}_i \tilde{\boldsymbol{\alpha}}) \tag{17}$$
if $\mathbf{x}_i$ is observed in area $i$ but $\bar{y}_i$ is not;
$$\tilde{\theta}_{iMFH} = \kappa_i \bar{y}_i + (1 - \kappa_i) [\mathbf{0}^T, 1] \mathbf{T}_i \tilde{\boldsymbol{\alpha}} \tag{18}$$
if $\bar{y}_i$ is observed but $\mathbf{x}_i$ is not.

The weighting $\kappa_i$ in the small area estimator in (16) to (18) depends on the variability of $\mathbf{x}_i$ as well as on the sampling variability of $\bar{y}_i$: $\kappa_i$ is smaller, and the small area estimator depends more heavily on the direct estimator, if the variability of $\mathbf{x}_i$ is larger. If $\mathbf{X}_i$ is measured exactly (i.e., all entries of $\boldsymbol{\Sigma}_i$ are 0), then $\tilde{\theta}_{iMFH}$, using assumptions of normality, coincides with the univariate Fay-Herriot estimator that incorporates the $\mathbf{X}_i$'s as covariates.

The MSE of the estimator in (16) to (18) can be obtained using standard methods and is given in Lohr and Ybarra (2003). As occurred with the unit-level model, use of the multivariate Fay-Herriot model results in improved efficiency.

In practice, $\boldsymbol{\Sigma}_b$ as well as $\boldsymbol{\alpha}$ must be estimated from the data. Method of moments, maximum likelihood, or restricted maximum likelihood may be used. See Datta et al. (2001) for a comparison of the estimators of $\boldsymbol{\Sigma}_b$ in the univariate case.

### 3.3. Measurement Error Model

As shown in Section 3.1, ignoring the error in $\mathbf{x}_i$ gives a biased mean squared error and a non-optimal weighting of the direct and indirect estimators. The motivation for using a measurement error model comes from the observation that omitted or inaccurate covariates can cause bias. Suppose that the model in (10) holds, but it is fitted omitting the term $\mathbf{X}_i^T\boldsymbol{\beta}$. Then estimates of the regression parameters $\boldsymbol{\alpha}$ and the predicted values may be biased. This bias leads to an increase in the MSE of the predicted values. If $\mathbf{x}_i$ or another estimator $\hat{\mathbf{X}}_i$ is included in the covariates, however, the error in measuring $\mathbf{X}_i$ must be accounted for in the estimation and mean squared error. Fuller (1987, 1990), Carroll et al. (1995) and Cheng and Van Ness (1999) discussed measurement error models for estimation of regression parameters and for prediction.

As before, let $\hat{\mathbf{X}}_i$ be an estimator of the population quantity $\mathbf{X}_i$ with $\text{MSE}\,(\hat{\mathbf{X}}_i) = \mathbf{C}_i$. We assume that such an estimator exists for every area: If $\mathbf{x}$ is not measured in area $i$, then an empirical Bayes estimator or imputed value may be used for $\hat{\mathbf{X}}_i$. Consider the model

$$\bar{y}_i = \mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta} + r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) + e_i \tag{19}$$

where

$$r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) = v_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)^T\boldsymbol{\beta}.$$

Here, $v_i \sim N(0, \sigma_v^2)$ represents the model error and $e_i \sim N(0, \psi_i)$ represents the design-based survey error for $\bar{y}_i$. We assume that $v_i$ is independent of both $\hat{\mathbf{X}}_i$ and $\bar{y}_i$. For simplicity, we also assume here that all $\hat{\mathbf{X}}_i$'s and $\bar{y}_i$'s are independent; Ybarra (2003) develops theory for the more general case. Consequently, $\text{MSE}(r_i) = \sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$. Now let

$$\tilde{\theta}_{iME} = \gamma_i\bar{y}_i + (1-\gamma_i)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}), \tag{20}$$

where

$$\gamma_i = \frac{\sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}}{\sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta} + \psi_i}. \tag{21}$$

Then $\tilde{\theta}_{iME}$ has minimum mean squared error among all linear combinations of $\bar{y}_i$ and $\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta}$ of the form $a_i\bar{y}_i + (1-a_i)(\mathbf{A}_i^T\boldsymbol{\alpha} + \hat{\mathbf{X}}_i^T\boldsymbol{\beta})$ where $0 \leq a_i \leq 1$. The estimator in (21) may also be derived as the "best" estimator in the Rao-Blackwell sense if normality is assumed.

The relative weights $\gamma_i$ depend on the error in estimating $\mathbf{X}_i$: $\gamma_i$ is smaller when $\hat{\mathbf{X}}_i$ is measured without error. If $\hat{\mathbf{X}}_i$ is measured imprecisely, then $\gamma_i$ is larger and the estimator depends more heavily on the direct estimator $\bar{y}_i$. If $\bar{y}_i$ is measured in area $i$ then $\text{MSE}(\tilde{\theta}_{iME}) = \psi_i\gamma_i$, which is at most as large as the variance $\psi_i$ of the direct estimator, $\bar{y}_i$. If $\bar{y}_i$ is not measured in area $i$ then $\text{MSE}\,(\tilde{\theta}_{iME}) = \sigma_v^2 + \boldsymbol{\beta}^T\mathbf{C}_i\boldsymbol{\beta}$.

Note that $\text{MSE}\,(\tilde{\theta}_{iME}) \leq \text{MSE}\,(\tilde{\theta}_i^*)$ where $\tilde{\theta}_i^*$ is the substitution estimator from (13): the two MSE's are equal if $\mathbf{C}_i = 0$. If the empirical Bayes estimator is used for $\hat{\mathbf{X}}_i$, then it can be shown that the estimator in (20) is equivalent to the multivariate Fay-Herriot estimator.

In practice, the quantities $\sigma_v^2$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown and must be estimated from the data. Lindley (1947, p. 243) suggested using weighted least squares to estimate the regression parameters. For our model, the MSE of the errors $(r_i + e_i)$ is $\psi_i + \sigma_v^2 + \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}$. Thus, one can solve for the unknown parameters by minimizing

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{(\bar{y}_i - \mathbf{A}_i^T \boldsymbol{\alpha} - \hat{\mathbf{X}}_i^T \boldsymbol{\beta})^2}{\psi_i + \sigma_v^2 + \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}}$$

where the sum is over areas $i$ where $\bar{y}_i$ is measured. Gleser (1981) gave large sample properties of the resulting estimates of the regression parameters. If $\sigma_v^2$ is unknown, we can use modified least squares to estimate the parameters (Cheng and Van Ness, 1999, pp. 85 and 146). In this case an unbiased estimator of $\sigma_v^2$ is

$$Q_1(\boldsymbol{\alpha}, \boldsymbol{\beta}, \psi_1, \ldots, \psi_m) = m^{-1} \sum_{i=1}^{m} [(\bar{y}_i - \mathbf{A}_i^T \boldsymbol{\alpha} - \hat{\mathbf{X}}_i^T \boldsymbol{\beta})^2 - \psi_i - \boldsymbol{\beta}^T \mathbf{C}_i \boldsymbol{\beta}] \qquad (22)$$

Minimizing $Q_2$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ gives estimates of the regression parameters. Note, though, that terms in (22) may be negative and it is possible that minimization will occur on the boundaries of the parameter space. Ybarra (2003) modified the estimators so that the expected values of the regression parameters are finite and derived properties of the models using these estimators. She also explored effects of estimating the variances from the data.

Although in some situations the measurement error model and multivariate Fay-Herriot method give similar results, we prefer the measurement error model for many practical situations. It is more flexible for choice of estimator of $\mathbf{X}_i$. In addition, robust methods may be used for estimating the regression parameters and variance terms, so that the measurement error model is adaptable for situations in which some of the $\mathbf{x}_i$'s are outliers due to variable quality of the data sources.

### 3.4. Applications

The measurement error model has an advantage over the multivariate Fay-Herriot approach in that means and variances of the auxiliary information can be estimated separately from the quantities from the primary survey. Missing values may be imputed, and imputation variance used for the MSE of $\hat{\mathbf{X}}_i$. This approach would work better than the multivariate Fay-Herriot approach for estimating victimization rates at the state level, using the Uniform Crime Reports (UCR) data as auxiliary information.

The UCR data sets give crimes reported each month by each of the approximately 19,000 law enforcement agencies in the United States. In a typical year, however, approximately 1/3 of the total possible month/agency cells are missing. If complete records only are used as auxiliary information in a Fay-Herriot-type model, the resulting small area estimates may be biased and will have standard errors that are too small because they do not account for the uncertainty in the auxiliary information. The multivariate Fay-Herriot approach can reduce some of this bias by incorporating administrative covariates to improve prediction of the UCR (essentially, including the imputation in the model). But the imputation will be done at the state level for annual data; this will not be as good as an imputation done separately using partial agency information and longitudinal trends with the monthly data.

Schalk (2003) studied imputation methods for the western region of the Uniform Crime Reports data. She evaluated the currently used hot deck method, nearest neighbor, and several regression models for imputing missing cells and found that the hot deck method is the least accurate. All of the models studied can give standard errors for the statewide quantities by using bootstrap or multiple imputation. Thus, by doing the imputation separately, the auxiliary information is more accurate and is accompanied by an estimate of precision $\mathbf{C}_i$ that can be used with the measurement error model for estimating victimization rates with NCVS data. With the imputed values from the UCR, we are now in a position to apply the measurement error models in Section 3.3 to obtain more accurate estimates of local victimization rates.

We are also currently using the models discussed in this section to obtain small area estimates of the prevalence of diabetes for 50 demographic subgroups based on race/ethnicity, gender, and age. In NHANES, diabetes prevalence is estimated using medical exams of plasma glucose levels, while in NHIS diabetes-related problems are assessed using the results of questionnaires. Correlation between the items in the two surveys is about 0.4; using the NHIS data as auxiliary information reduces the MSE for diabetes prevalence in small demographic groups (with NHANES sample sizes between 5 and 7) by 40-80%.

## 4. Multiple Frame Surveys for Small Area Estimation

Up to this point, we have discussed using a second survey to provide auxiliary information for estimating a quantity of interest measured in the primary survey. The models given in Sections 2 and 3 use all available information for predicting $\theta_i$; if area $i$ has no observations from either the primary or secondary survey, then $\hat{\theta}_i$ relies on the predicted value from the regression using the administrative data. This may be the best that can be done with the available information, but sometimes a different design can give more precision for the direct estimators and for the estimated regression parameters.

One such design that can be used is a multiple frame survey. In a multiple frame survey, probability samples are drawn independently from $Q$ frames $A_1, \ldots, A_Q$. The union of the $Q$ frames is assumed to cover the finite population of interest, $\mathcal{U}$. The frames may overlap, resulting in a possible $2^Q - 1$ nonoverlapping domains.

Rao (2003, chapter 2) discussed the use of multiple frame designs for improving small area estimation. The primary purpose of many surveys is estimation of quantities such as unemployment or criminal victimization at the national level; the designs for the surveys thus are directed toward the national estimates, even though some surveys contain design features useful for small area estimation. These surveys, though, can be supplemented with additional samples from small areas of interest, so that the original survey and additional samples can be considered as a multiple frame survey. Madans et al. (2001) discussed using multiple frame surveys for supplementing information from NHIS; additional surveys may be taken from different states and combined with NHIS data for improved estimation at the state level. In this situation the same questions may be used in NHIS and the supplementary surveys.

Various estimators that have been proposed for combining information from the separate samples were reviewed in Lohr and Rao (2000). These estimators modify the weights associated with sampled units from each frame, so that the overall population total is estimated by a weighted sum of the observations from all of the samples using the modified weights. Many of these methods, however, were developed for estimating one population total or mean at a time, and use a different set of modified weights for each characteristic of interest. Such an approach will give nearly optimal results for individual responses, but will not work well for estimating small area totals or means directly from the surveys: If different weights are used for estimating the population total in different small areas, the sum of estimated small area population totals will not equal the estimated total for larger areas. It is thus desirable to have methods for obtaining direct small area estimates from multiple frame surveys that use the same set of weights for all variables. Skinner and Rao (1996) developed a pseudo-maximum likelihood method that uses the same weights for all variables for the two-frame situation.

Lohr and Rao (2002) developed estimation methods for multiple frame surveys with more than two frames that use the same weights for each variable being estimated, and thus can be applied when supplemental surveys are taken in several small areas. These methods easily apply to the small area setting by letting the variable of interest be the value $\theta_i$ for the $i^{th}$ small area. The improved direct estimators of the $\theta_i$'s may then be used with an area-level model to achieve greater efficiency.

## 5. Discussion and Future Work

In this paper, we have summarized recent research we have done on combining information from different sources for small area estimation. In many situations, much greater efficiency can be achieved by using auxiliary information from another survey. We believe that these methods have the potential to increase the accuracy of small area estimates with no or minimal increase in the cost of data collection, as they are all based on more efficient use of existing data.

The American Community Survey is intended, through its large sample size, to provide improved direct small area estimates for income and poverty. Those characteristics and other quantities measured in the ACS can also provide valuable and timely auxiliary data for small area estimation of quantities measured in other surveys. The methods summarized in this paper can be used to take advantage of this new, detailed data source for small area estimation of many different characteristics of interest.

Since the ACS uses rolling samples, longitudinal methods will also be helpful when using the ACS as auxiliary information. We are currently working on incorporating time series models into the estimation, and on obtaining longitudinal estimates from multiple frame surveys. A related problem is using spatial models to better include geographic information.

Another important problem under study is robustness to the model and to methods for estimating model quantities. One challenge of using UCR data as auxiliary information for the NCVS, in addition to the missing values, is that some agencies provide inaccurate estimation. These inaccuracies could then bias the results. Using robust methods is expected to reduce the effects of possible UCR outliers and result in more accurate small area estimates.

## Acknowledgements

This research was supported by the National Science Foundation through Grant SES 0105852. We are grateful for this support, and for helpful comments from J.N.K. Rao, John Eltinge, and the late Charles Alexander.

## References

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

Botman, S.L., Moore, T.F., Moriarity, C.L. and Parsons, V.L. (2000). "Design and Estimation for the National Health Interview Survey, 1995-2004," National Center for Health Statistics. *Vital Health Statistics* 2(130).

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models,* London: Chapman & Hall.

Cheng, C.-L. and Van Ness, J.W. (1999). *Statistical Regression with Measurement Error.* London: Arnold.

Datta, G. A., B. Day and I. Basawa (1999). "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation," *Journal of Statistical Planning and Inference*, 75, 269–279.

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

Datta, G. S., J. N. K. Rao and Smith, D. D. (2001), "On Measures of Uncertainty of Small Area Estimators in the Fay-Herriot Model," Technical Report, University of Georgia, Athens, Georgia.

Dueck, A. C. and Lohr, S. L. (2003), "Robust Estimation of Multivariate Covariance Components," manuscript submitted for publication.

Fay, R.E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in R. Platek et al. (eds.), *Small Area Statistics,* New York: Wiley, 91-102.

Fay, R. E. and R. A. Herriot (1979), "Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 78, 269-277.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Fuller, W. A. (1990), "Prediction of True Values for the Measurement Error Model," in P. J. Brown and W. A. Fuller (eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics Vol. 112,* Providence, RI: AMS, 41-57.

Gleser, L. J. (1981), "Estimation in a Multivariate 'Errors-in-Variables' Regression Model: Large Sample Results," *Annals of Statistics,* 9, 24–44.

Lindley, D.V. (1947), "Regression Lines and the Linear Functional Relationship," *Journal of the Royal Statistical Society Supp.,* 9, 218–244.

Lohr, S. (2003), "Privacy and Survey Research: Ethical and Legal Questions from a Researcher's Perspective," *Bulletin of the International Statistical Institute, Proceedings of the 54th Session.*

Lohr, S. and N. G. N. Prasad (2003), "Small Area Estimation with Auxiliary Survey Data," to appear in *Canadian Journal of Statistics.*

Lohr, S. and J. N. K. Rao (2000), "Inference in Dual Frame Surveys," *Journal of the American Statistical Association,* 95, 271–280.

Lohr, S. and J. N. K. Rao (2002), "Estimation in Multiple Frame Surveys," to appear in *Proceedings of the International Conference on Recent Advances in Survey Sampling.*

Lohr, S. and L. M. R. Ybarra (2003), "Area-level Models using Data from Multiple Surveys," to appear in *Proceedings of Statistics Canada Symposium 2002.*

Madans, J. H., T. M. Ezzati-Rice, M. Cynamon and S. J. Blumberg (2001), "Targeting Approaches to State-Level Estimates," in M.L. Cynamon and R.A. Kulka (eds.) *Proceedings of the Seventh Conference on Health Survey Research Methods*, Hyattsville, MD, Department of Health and Human Services, 239–245.

Maltz, M. (1999), *Bridging Gaps in Police Crime Data*, NCJ Report 176365, Washington, D.C.: Bureau of Justice Statistics.

National Research Council (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond.* C. F. Citro and R. T. Michael, eds. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

Rao, J. N. K. (2003), *Small Area Estimation,* New York: Wiley.

Schalk, K. L. (2003), *Imputation of Missing Uniform Crime Report Data,* unpublished M.S. thesis, Arizona State University.

Skinner, C. J. and J.N.K. Rao (1996). "Estimation in Dual Frame Surveys With Complex Designs," *Journal of the American Statistical Association*, 91, 349–356.

Wiersema, B., McDowall, D. and Loftin, C. (2000), “Comparing Metropolitan Area Estimates of Crime from the National Crime Victimization Survey and Uniform Crime Reports,” Paper presented at the Joint Statistical Meetings, Indianapolis, August 2000.

Ybarra, L. M. R. (2003), *Area-level Models Using Data from Multiple Surveys*, unpublished Ph.D. dissertation, Arizona State University.

Ybarra, L. M. R. and Lohr, S. (2002), “Estimates of Repeat Victimization Using the National Crime Victimization Survey,” *Journal of Quantitative Criminology,* 18, 1–21.