Survey Methodology, June 2015
Vol. 41, No. 1, pp. 21-36
Statistics Canada, Catalogue No. 12-001-X

Small area estimation combining information from several sources

Jae-kwang Kim, Seunghwan Park and Seo-young Kim[1]

Abstract

An area-level model approach to combining information from several sources is considered in the context of small area estimation. At each small area, several estimates are computed and linked through a system of structural error models. The best linear unbiased predictor of the small area parameter can be computed by the general least squares method. Parameters in the structural error models are estimated using the theory of measurement error models. Estimation of mean squared errors is also discussed. The proposed method is applied to the real problem of labor force surveys in Korea.

Key Words: Area-level model; Auxiliary information; Measurement error models; Structural error model; Survey integration.

1 Introduction

Combining information from different sources is an important problem in statistics. In survey sampling, combining information from multiple surveys can improve the quality of small area estimates. The source of information can come from a probability sample with direct measurements, from another probability sample with indirect measurements (such as self-reported health status), or from auxiliary area-level information. Many approaches of combining information, such as the multiple-frame and statistical matching methods, require access to individual level data, which is not always feasible in practice.

We consider an area-level model approach to small area estimation when there are several sources of auxiliary information. Pfeffermann (2002) and Rao (2003) provided thorough reviews of methods used in small area estimation. Lohr and Prasad (2003) used multivariate models to combine information from several surveys. Ybarra and Lohr (2008) considered the small area estimation problem when the area-level auxiliary information has measurement errors. Merkouris (2010) discussed the small area estimation by combining information from multiple surveys. Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer (2007) and Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011) used Bayesian hierarchical models to combine information from multiple surveys for small area estimation. Kim and Rao (2012) considered a design-based approach to combining information from two independent surveys.

To describe the setup, suppose that the finite population consists of H subpopulations, denoted by $U_1, \ldots, U_H$, and that we are interested in estimating the subpopulation totals $X_h = \sum_{i \in U_h} x_i$ of a variable x for each area h. We assume that there is a survey that measures x_i from the sample but its sample size is not large enough to obtain estimates for X_h with reasonable accuracy. Consider one of the surveys, called survey A, as the main survey, and let $\hat{X}_h$ denote a design-consistent estimator of X_h obtained

1. Jae-kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.; Seunghwan Park, Department of Statistics, Seoul National University, Seoul, 151-747, Korea. E-mail: kkampsh@gmail.com; Seo-young Kim, Statistical Research Institute, Statistics Korea, Daejon, 302-847, Korea.

from survey A. Often, we compute $\hat{X}_h = \sum_{i \in A_h} w_{ia} x_i$, where A_h is the set of sample A for subpopulation h and w_{ia} is the weight of unit i in sample A.

In addition to the main survey, suppose that there is another survey, called survey B, that measures a rough estimate for x_i. Let y_{1i} be the measurement taken from survey B. We may assume that y_{1i} is a rough measurement of x_i with some level of measurement error. Thus, we may assume

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{1.1}$$

for some (β_0, β_1), where $e_{1i} \sim (0, \sigma_{e1}^2)$. Model (1.1) is variable-specific and the linear regression assumption or equal variance assumptions can be relaxed later. If $(\beta_0, \beta_1) = (0,1)$, then model (1.1) means that there is no measurement bias. Note that model parameters (β_0, β_1) in (1.1) are not area specific, but may be different for groups of areas, as demonstrated in the Korean labor force survey application in Section 5. Separate regression models for different groups may lead to smaller model errors and thus improve the statistical efficiency of the proposed method. From survey B, we can obtain another estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ of X_h, where w_{ib} is the weight of unit i in the sample from survey B and B_h is the B-sample for subpopulation h. Note that $\hat{Y}_{1h}$ can be obtained, for each area, if the same areas are identified in both surveys A and B. Model (1.1) can be used to combine information from the two surveys.

Finally, another source of information can be the Census information. Census information does not suffer from coverage error or sampling error. But, it may have measurement errors and it does not provide updated information for each month or year. Let y_{2i} be the measurement for unit i from the Census. The subpopulation total $Y_{2h} = \sum_{i \in C_h} y_{2i}$ is available when C_h is the set of Census C for subpopulation h.

Table 1.1 summarizes the major sources of information that we can consider into small area estimation.

Table 1.1
Available information for small area estimation

Data	Observation	Area level estimate	Properties
Survey A	direct obs. (x_i)	$\hat{X}_h, \hat{V}(\hat{X}_h)$	Sampling error (large)
Survey B	aux. obs. (y_{1i})	$\hat{Y}_{1h}, \hat{V}(\hat{Y}_{1h})$	Bias Measurement error Sampling error
Census	aux. obs. (y_{2i})	Y_{2h}	Measurement error No updated information

In this paper, we consider an area-level model approach for small area estimation combining all available information. The proposed approach is based on the measurement error models, where the sampling errors of the direct estimators are treated as measurement errors, and all the other auxiliary information are combined through a set of linking models. The proposed approach is applied to the small area estimation problem for labor force surveys in Korea, where three estimates are combined to produce small area estimates for unemployment rates.

The paper is organized as follows. In Section 2, the basic setup is introduced and the small area estimation problem is viewed as a measurement error model prediction problem. In Section 3, parameter estimation for the area level small area model is discussed. In Section 4, estimation of mean squared error is briefly discussed. In Section 5, the proposed method is applied to the labor force survey data in Korea. Concluding remarks are made in Section 6.

2 Basic theory

In this section, we first introduce the basic theory for combining the information for small area estimation. We first consider the simple case of combining two surveys. Assume that there are two surveys, survey A and survey B, obtained from separate probability sampling designs. The two surveys are not necessarily independent. From survey A, we obtain a design unbiased estimator $\hat{X}_{h,a} = \sum_{i \in A_h} w_{ia} x_i$ and its variance estimator $\hat{V}\left(\hat{X}_h\right)$. From survey B, we obtain a design unbiased estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ of $Y_{1h} = \sum_{i \in U_h} y_{1i}$. The sampling error of $\left(\hat{X}_h, \hat{Y}_{1h}\right)$ can be expressed by the *sampling error model*

$$\begin{pmatrix} \hat{X}_h \\ \hat{Y}_{1h} \end{pmatrix} = \begin{pmatrix} X_h \\ Y_{1h} \end{pmatrix} + \begin{pmatrix} N_h a_h \\ N_h b_h \end{pmatrix} \tag{2.1}$$

and a_h and b_h represent the sampling errors associated with $\hat{X}_h / N_h$ and $\hat{Y}_{1h} / N_h$ such that

$$\begin{pmatrix} a_h \\ b_h \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V(a_h) & \text{Cov}(a_h, b_h) \\ \text{Cov}(a_h, b_h) & V(b_h) \end{pmatrix} \right].$$

Our parameter of interest is the population total X_h of x in area h.

From (1.1), we obtain the following area level model:

$$Y_{1h} = N_h \beta_0 + \beta_1 X_h + \tilde{e}_{1h}, \tag{2.2}$$

where $(N_h, X_h, Y_{1h}, \tilde{e}_{1h}) = \sum_{i \in U_h} (1, x_i, y_{1i}, e_{1i})$. We can express (2.2) in terms of population mean

$$\bar{Y}_{1h} = \beta_0 + \bar{X}_h \beta_1 + \bar{e}_{1h}, \tag{2.3}$$

where $(\bar{X}_h, \bar{Y}_{1h}, \bar{e}_{1h}) = N_h^{-1} \sum_{i \in U_h} (x_i, y_{1i}, e_{1i})$. If we use a nested error model

$$e_{1hi} = \varepsilon_h + u_{hi} \tag{2.4}$$

where $\varepsilon_h \sim \left(0, \sigma_e^2\right)$ and $u_{hi} \sim \left(0, \sigma_u^2\right)$, then $\bar{e}_{1h} \sim \left(0, \sigma_{e,h}^2\right)$, $\sigma_{e,h}^2 = \sigma_e^2 + \sigma_u^2 / N_h$. The nested error model is quite popular in small area estimation (e.g., Battese, Harter and Fuller 1988) and it assumes that $\text{Cov}\left(e_{1hi}, e_{1hj}\right) = \sigma_e^2$ for $i \neq j$. Because N_h is often quite large, we can safely assume that $\bar{e}_{1h} \sim \left(0, \sigma_{e,h}^2 = \sigma_e^2\right)$. The model (2.2) is called *structural error model* because it describes the structural

relationship between the two latent variables Y_{1h} and X_h. The two models, (2.1) and (2.2), are often encountered in the measurement error model literature (Fuller 1987). Thus, the model for small area estimation can be viewed as a measurement error model, as suggested by Fuller (1991) who originally used the measurement error model approach in the unit-level modeling for small area estimation.

Now, if we define $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1}\left(\hat{Y}_{1h}, \hat{X}_h\right)$, combining (2.1) and (2.3), we have

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

which can also be written as

$$\begin{pmatrix} \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}. \tag{2.5}$$

Thus, when all the model parameters in (2.5) are known, the best estimator of $\bar{X}_h$ can be computed by

$$\hat{\bar{X}}_h = \left\{(\beta_1, 1) V_h^{-1} (\beta_1, 1)'\right\}^{-1} (\beta_1, 1) V_h^{-1} (\bar{y}_{1h} - \beta_0, \bar{x}_h)' \tag{2.6}$$

where V_h is the variance-covariance matrix of $(b_h + \bar{e}_{1h}, a_h)'$. The variance of $\hat{\bar{X}}_h$ is given by $\left\{(\beta_1, 1) V_h^{-1} (\beta_1, 1)'\right\}^{-1}$. The estimator in (2.6) can be called the Generalized Least Squares (GLS) estimator because it uses the technique of the generalized least squares method in the linear model theory. The GLS method is useful because it is optimal and it can incorporate additional sources of information naturally. For example, if another estimator $\bar{y}_{2h}$ for $\bar{Y}_{2h}$ is also available and satisfies

$$\bar{Y}_{2h} = \gamma_0 + \gamma_1 \bar{X}_h + \bar{e}_{2h}$$

and

$$\bar{y}_{2h} = \bar{Y}_{2h} + c_h,$$

then the extended GLS model is written as

$$\begin{pmatrix} \bar{y}_{2h} - \gamma_0 \\ \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} c_h + \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \tag{2.7}$$

and the GLS estimator can be obtained by

$$\hat{\bar{X}}_{h2} = \left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1} (\gamma_1, \beta_1, 1) V_{h2}^{-1} (\bar{y}_{2h} - \gamma_0, \bar{y}_{1h} - \beta_0, \bar{x}_h)'$$

where V_{h2} is the variance-covariance matrix of $(c_h + \bar{e}_{2h}, b_h + \bar{e}_{1h}, a_h)'$. The GLS estimator has variance $\left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1}$. If $\bar{y}_{2h}$ is independent of $(\bar{x}_h, \bar{y}_{1h})$, the efficiency gain by incorporating $\bar{y}_{2h}$ into GLS in terms of relative variance can be expressed as

$$\frac{V\left(\hat{\bar{X}}_{h2}\right) - V\left(\hat{\bar{X}}_{h}\right)}{V\left(\hat{\bar{X}}_{h}\right)} = -\frac{\{V(\bar{y}_{2h}/\gamma_1)\}^{-1}}{\left\{V\left(\hat{\bar{X}}_{h}\right)\right\}^{-1} + \{V(\bar{y}_{2h}/\gamma_1)\}^{-1}},$$

where $V(\bar{y}_{2h}/\gamma_1) = V(c_h + \bar{e}_{2h})/\gamma_1^2$. The gain is high if both the sampling variance of $\bar{y}_{2h}$ and the model variance $V(\bar{e}_{2h})$ are small. If $\gamma_1 = 0$, then there is no gain.

Remark 1 *Note that model (2.5) can also be written as*

$$\begin{pmatrix} \beta_1^{-1}(\bar{y}_{1h} - \beta_0) \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} (b_h + \bar{e}_{1h})/\beta_1 \\ a_h \end{pmatrix}. \tag{2.8}$$

The GLS estimator obtained from (2.8), which is the same as the GLS estimator obtained from (2.5), can be expressed as

$$\hat{\bar{X}}_h = \alpha_h \bar{x}_h + (1 - \alpha_h)\tilde{x}_h \tag{2.9}$$

where $\tilde{x}_h = \beta_1^{-1}(\bar{y}_{1h} - \beta_0)$ *and*

$$\begin{aligned} \alpha_h &= \frac{V(\tilde{x}_h) - \mathrm{Cov}(\bar{x}_h, \tilde{x}_h)}{V(\bar{x}_h) + V(\tilde{x}_h) - 2\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)} \\ &= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 \mathrm{Cov}(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 \mathrm{Cov}(a_h, b_h)}. \end{aligned}$$

The estimator $\tilde{x}_h$, *when computed with estimated parameter* $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right)$, *is called the synthetic estimator and the optimal estimator in (2.9) is often called the composite estimator. It can be shown that, ignoring the effect of estimating* β, *the variance of the composite estimator is equal to*

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{2.10}$$

and, as $\alpha_h < 1$, *the composite estimator is more efficient than the direct estimator.*

3 Parameter estimation

Now, we discuss estimation of the model parameters in (2.3). The GLS estimator of $\beta = (\beta_0, \beta_1)$ can be obtained by minimizing

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{3.1}$$

Since

$$V(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1) = \sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)', \tag{3.2}$$

where $\sigma_{e,h}^2 = V(\bar{e}_{1h})$ and $\Sigma_h = V\left\{(a_h, b_h)'\right\}$, we can express

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} w_h(\beta_1)(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2, \tag{3.3}$$

where $w_h(\beta_1) = \left\{\sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)'\right\}^{-1}$. Now, by solving $\partial Q^* / \partial \beta = 0$, we have

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \tag{3.4}$$

and

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\{(\bar{x}_h - \bar{x}_w)(\bar{y}_{1h} - \bar{y}_{1w}) - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{(\bar{x}_h - \bar{x}_w)^2 - V(a_h)\right\}}, \tag{3.5}$$

where

$$(\bar{x}_w, \bar{y}_w) = \left\{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\right\}^{-1} \sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)(\bar{x}_h, \bar{y}_h).$$

Note that the weight $w_h(\beta_1)$ depends on β_1. Thus, the solution (3.5) can be obtained by an iterative algorithm. Once $\hat{\beta}_1$ is computed by (3.5), then $\hat{\beta}_0$ is obtained by (3.4).

Now, we discuss the estimation of model variance $\sigma_{e,h}^2$. The simplest method is the Method of Moments (MOM). That is, we can use

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_{e,h}^2 \tag{3.6}$$

to obtain an unbiased estimator of $\sigma_{e,h}^2$. Under the nested error model in (2.4), we have $\sigma_{e,h}^2 = \sigma_e^2$ and

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_e^2. \tag{3.7}$$

Thus, similarly to Fuller (2009), the MOM estimator of σ_e^2 can be obtained by

$$\hat{\sigma}_e^2 = \sum_{h=1}^{H} \kappa_h \left\{\left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h\hat{\beta}_1\right)^2 - \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\} \tag{3.8}$$

where

$$\kappa_h \propto \left\{\hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

and $\sum_{h=1}^{H} \kappa_h = 1$. Because κ_h depends on $\hat{\sigma}_e^2$, the solution (3.8) can be obtained iteratively, using $\hat{\sigma}_e^2 = 0$ as an initial value. Fay and Herriot (1979) used an alternative method which is based on the iterative solution to nonlinear equation:

$$\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h\right)^2}{\sigma_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)'} = H - 2.$$

Writing the above equation as $g\left(\sigma_e^2\right) = H - 2$, a Newton-type method for $g(\theta) = 0$ with $\theta = \sigma_e^2$ can be obtained by

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{g'\left(\theta^{(t)}\right)}\left(H - 2 - g\left(\theta^{(t)}\right)\right) \tag{3.9}$$

where

$$g'(\theta) = -\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h\right)^2}{\left\{\theta + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)'\right\}^2}.$$

Assuming $\sigma_{e,h}^2 \equiv \sigma_e^2$, we now describe the whole parameter estimation procedure as follows:

Step 1 Compute the initial estimator of (β_0, β_1) by setting $\hat{\sigma}_e^2 = 0$ in (3.4) and (3.5).

Step 2 Based on the current value of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, compute $\hat{\sigma}_e^2$ using the iterative algorithm in (3.9).

Step 3 Use the current value of $\hat{\sigma}_e^2$, compute the updated estimator of (β_0, β_1) by (3.4) and (3.5).

Step 4 Repeat [Step 2]-[Step 3] until convergence.

The proposed parameter estimation method estimates $\beta = (\beta_0, \beta_1)$ by the GLS and estimates σ_e^2 by the MOM iteratively. Note that the estimation of β is based on data from all areas. If separate regression models are used, then the proposed parameter estimation method can be applied to the groups of areas. Instead of this separate iterative estimation method, we can also consider another method based on maximum likelihood estimation (MLE) under parametric distributional assumptions. See Carroll, Rupert, and Stefanski (1995) and Schafer (2001) for further discussion of MLE for parameters in the measurement error models.

Remark 2 *If $\sigma_{e,h}^2 = \sigma_e^2$ is not true, we can consider some alternative model such as*

$$\bar{e}_h \sim \left(0, \bar{X}_h \sigma_e^2\right). \tag{3.10}$$

To check whether model (3.10) holds, one can compute

$$\nu_h = \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1\right)^2 - \hat{\beta}_1^2 V(a_h) + 2\hat{\beta}_1 \hat{C}(a_h, b_h) - V(b_h) \tag{3.11}$$

and plot ν_h on $\bar{x}_h$. If the plot shows a linear relationship, then (3.10) can be treated as a reasonable model. Under model (3.10), we can obtain σ_e^2 by a ratio method:

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^{H} \kappa_h \nu_h}{\sum_{h=1}^{H} \kappa_h \hat{\bar{X}}_h} \tag{3.12}$$

where

$$\kappa_h \propto \left\{ \hat{\bar{X}}_h \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right) \right\}^{-1}$$

with $\sum_{h=1}^{H} \kappa_h = 1$, $\hat{\bar{X}}_h$ *is defined in (2.9), and* ν_h *is defined in (3.11). Because* κ_h *also depends on* σ_e^2, *the solution (3.12) can be obtained iteratively.*

Remark 3 *We can also consider a transformation* $\bar{x}_h^* = T(\bar{x}_h)$ *and* $\bar{y}_{1h}^* = T(\bar{y}_{1h})$ *to improve the approximation to asymptotic normality. To check the departure from normality, plot* $n_{ha} \bar{V}(\bar{x}_h)$ *on* $\bar{x}_h$. *If the plot shows some structural relationship of* $\bar{x}_h$ *then the normality assumption can be doubted. Now, consider the following transformation*

$$T(x) = \log(x). \tag{3.13}$$

Note that the asymptotic variance of $\bar{x}_h^* = T(\bar{x}_h)$ *is equal to*

$$V\left(\bar{x}_h^*\right) \doteq \frac{1}{\left(\bar{x}_h\right)^2} V(\bar{x}_h).$$

Such transformation is a variance stabilizing transformation and is useful when we want to improve the approximation to normality.

Once the GLS estimator $\hat{\bar{X}}_h^*$ *of* $\bar{X}_h^*$ *is obtained, then we need to apply the inverse transformation to obtain the best estimator of* $\bar{X}_h = T^{-1}\left(\bar{X}_h^*\right) := Q\left(\bar{X}_h^*\right)$. *Simply applying the inverse transformation will lead to biased estimation. To correct for the bias, we can use a second-order Taylor linearization. Using a Taylor expansion, we have*

$$Q\left(\hat{\bar{X}}_h^*\right) \doteq Q\left(\bar{X}_h^*\right) + Q'\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right) + \frac{1}{2} Q''\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right)^2$$

and so, if we use $Q\left(\hat{\bar{X}}_h^*\right)$ *as an estimator for* $\bar{X}_h = Q\left(\bar{X}_h^*\right)$, *we have, ignoring the smaller order terms,*

$$E\left\{Q\left(\hat{\bar{X}}_h^*\right)\right\} = \bar{X}_h + \frac{1}{2} Q''\left(\bar{X}_h^*\right) V\left(\hat{\bar{X}}_h^*\right).$$

For the transformation in (3.13), we have $Q\left(\bar{X}_h^*\right) = \exp\left(\bar{X}_h^*\right)$ *and so* $Q''\left(\bar{X}_h^*\right) = \bar{X}_h$. *Thus,* $\hat{\bar{X}}_h = Q\left(\hat{\bar{X}}_h^*\right)$, *we have*

$$E\left(\hat{\bar{X}}_h\right) \cong \bar{X}_h + \frac{1}{2} \bar{X}_h V\left(\hat{\bar{X}}_h^*\right)$$

and the bias-corrected estimator of $\bar{X}_h$ *is*

$$\hat{\bar{X}}_{h,bc} = \frac{\hat{\bar{X}}_h}{1 + 0.5V\left(\hat{\bar{X}}_h^*\right)}, \tag{3.14}$$

where $V\left(\hat{\bar{X}}_h^*\right)$ *is computed by the MSE estimation method which will be discussed in Section 4.*

4 MSE estimation

We now discuss mean squared error (MSE) estimation of the GLS estimator $\hat{\bar{X}}_h$ which is given by (2.9). Note that the GLS estimator is a function of (β_0, β_1) and σ_e^2. If the model parameters are known, then the MSE of $\hat{\bar{X}}_h$ is equal to $M_{h1} = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)$, as discussed in Remark 1. That is, writing $\theta = \left(\beta_0, \beta_1, \sigma_e^2\right)$ and $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\theta)$, the actual prediction for $\bar{X}_h$ is computed by $\hat{\bar{X}}_{eh} = \hat{\bar{X}}_h\left(\hat{\theta}\right)$. To account for the effect of estimating the model parameters, we first note the following decomposition of $\mathrm{MSE}\left(\hat{\bar{X}}_h^*\right)$:

$$\begin{aligned} \mathrm{MSE}\left(\hat{\bar{X}}_{eh}\right) &= \mathrm{MSE}\left(\hat{\bar{X}}_h\right) + E\left\{\left(\hat{\bar{X}}_{eh} - \hat{\bar{X}}_h\right)^2\right\} \\ &=: M_{h1} + M_{h2}, \end{aligned}$$

which was originally proved by Kackar and Harville (1984) under normality assumptions. The first term, M_{h1}, is of order $1/n_h$, where n_h is the size of A_h, and the second term, M_{h2}, is of order $1/n$ with $n = \sum_{h=1}^{H} n_h$. The second term is often much smaller than the first term.

We consider a jackknife approach to estimate the MSE. Use of the jackknife for bias-corrected estimation was originally proposed by Quenouille (1956). Jiang, Lahiri and Wan (2002) provided a rigorous justification of the jackknife method for the MSE estimation in small area estimation. The following steps can be used for the jackknife computation.

Step 1 Calculate the k^{th} replicate $\hat{\theta}^{(-k)}$ of $\hat{\theta}$ by deleting the k^{th} area data set $(\bar{x}_k, \bar{y}_{1k})$ from the full data set $\{(\bar{x}_h, \bar{y}_{1h}); h = 1, 2, \ldots, H\}$. This calculation is done for each k to get H replicates of θ: $\left\{\hat{\theta}^{(-k)}; k = 1, \ldots, H\right\}$ which, in turn, provide H replicates of $\hat{\bar{X}}_h$: $\left\{\hat{\bar{X}}_h^{(-k)}; k = 1, 2, \ldots, H\right\}$, where $\hat{\bar{X}}_h^{(-k)} = \hat{\bar{X}}_h\left(\hat{\theta}^{(-k)}\right)$.

Step 2 Calculate the estimator of M_{h2} as

$$\hat{M}_{2h} = \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\bar{X}}_h^{(-k)} - \hat{\bar{X}}_h\right)^2. \tag{4.1}$$

Step 3 Calculate the estimator of M_{h1} as

$$\hat{M}_{1h} = \hat{\alpha}_h^{(\mathrm{JK})} V(\bar{x}_h) + \left(1 - \hat{\alpha}_h^{(\mathrm{JK})}\right)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{4.2}$$

where $\hat{\alpha}_h^{(\mathrm{JK})}$ is a bias-corrected estimator of α_h given by

$$\begin{aligned}\hat{\alpha}_h^{(\mathrm{JK})} &= \hat{\alpha}_h - \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\alpha}_h^{(-k)} - \hat{\alpha}_h\right),\\ \hat{\alpha}_h &= \frac{\hat{\sigma}_e^2 + V(b_h) - \hat{\beta}_1 \mathrm{Cov}(a_h, b_h)}{\hat{\sigma}_e^2 + V(b_h) + \hat{\beta}_1^2 V(a_h) - 2\hat{\beta}_1 \mathrm{Cov}(a_h, b_h)},\end{aligned}$$

and

$$\hat{\alpha}_h^{(-k)} = \frac{\hat{\sigma}_e^{(-k)2} + V(b_h) - \hat{\beta}_1^{(-k)} \mathrm{Cov}(a_h, b_h)}{\hat{\sigma}_e^{(-k)2} + V(b_h) + \left(\hat{\beta}_1^{(-k)}\right)^2 V(a_h) - 2\hat{\beta}_1^{(-k)} \mathrm{Cov}(a_h, b_h)}.$$

Remark 4 *For the transformation in (3.13), we use the bias-corrected estimator in (3.14) and its MSE estimation method needs to be changed. Using* $\hat{\bar{X}}_{eh,bc}$ *to denote the bias-corrected estimator in (3.14) evaluated at* $\hat{\theta}$*, we can have the*

$$\begin{aligned}\mathrm{MSE}\left(\hat{\bar{X}}_{eh,bc}\right) &= \mathrm{MSE}\left(\hat{\bar{X}}_{eh}\right)\\ &= \mathrm{MSE}\left\{Q\left(\hat{\bar{X}}_{eh}^*\right)\right\}\\ &\cong \left\{Q'\left(\bar{X}_h^*\right)\right\}^2 \cdot \mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right)\\ &= \bar{X}_h^2 \cdot \mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right),\end{aligned}$$

where the first equality follows that $\hat{\bar{X}}_{h,bc} - \hat{\bar{X}}_h$ *is of order* $O_p\left(n_h^{-1}\right)$*. The MSE of* $\hat{\bar{X}}_h^*$*, the EGLS estimator of* $\bar{X}_h^*$ *after transformation, is computed by (4.1) and (4.2). Once* $\mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right)$ *is estimated, we should multiply it by* $\hat{\bar{X}}_h^2$ *to obtain the MSE estimator of the back-transformed EGLS estimator* $\hat{\bar{X}}_{eh,bc}$*.*

5 Application to Korean Labor Force survey

We now consider an application of the proposed method to the labor force surveys in Korea. In Korea, two different labor force surveys are used to obtain information about employment. One is the Korean Labor Force (KLF) survey and the other is the Local Area labor force (LALF) survey. The KLF survey has about 7K sample households but LALF has about 200K sample households. Because LALF is a large-scale survey employing a lot of part time interviewers, there is a certain level of measurement errors in the LALF survey. We assume that the KLF has no measurement error, although it has significant sampling errors at the small area level. The KLF sample is a second-phase sample from the LALF sample. Thus, the sampling errors for two survey estimates are correlated. Let $\bar{X}_h$ be the (true) unemployment rate for area h. The small area level we considered is called “Gu”. The number of “Gu” in Korea is 229.

We observe $\bar{x}_h$ from KLF survey and $\bar{y}_{1h}$ from the LALF survey. To construct linking models, we first partition the population into two regions, urban region and rural region, based on the proportion of the households working on agricultural practice. Within each region, we build models separately (same model but allows for different parameter) and estimate the model parameters separately. The structural model is

$$\bar{Y}_h = \beta_1 \bar{X}_h + e_h \tag{5.1}$$

with $e_h \sim (0, \sigma_e^2)$. Here, we set $\beta_0 = 0$ to guarantee that the GLS estimator of $\bar{X}_h$ is nonnegative. The sampling error model remains the same. In this case, β_1 can be estimated by

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{\bar{x}_h \bar{y}_{1h} - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{\bar{x}_h^2 - V(a_h)\}}. \tag{5.2}$$

The sampling variance of (a_h, b_h) is computed using the method of reversed two-phase sampling described in the Appendix. The model variance is estimated by the method of moment technique in (3.8) with $\hat{\beta}_0 = 0$. The GLS estimator can be computed by (2.9) with $\tilde{x}_h = \hat{\beta}_1^{-1} \bar{y}_{1h}$.

In addition to the two surveys, we can also use the Census information. The GLS model incorporating the three sources of information can be expressed as

$$\begin{pmatrix} \bar{Y}_{2h} \\ \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

where $\bar{Y}_{2h}$ is the census result for area h. Because the Census estimate does not suffer from sampling error, we have only model error e_{2h} which represents the error when we model $E(\bar{Y}_{h2}) = \gamma_1 \bar{X}_h$. The model parameters can be obtained using the method in Section 3 with $\Sigma_h = \text{diag}(0, V(a_h, b_h))$. The GLS estimator of $\bar{X}_h$ can be obtained easily. The MSE part can be computed by using the fact that

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \right\}^{-1} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} := M_{h1}$$

and applying the jackknife method for bias correction.

Figure 5.1 presents the plot of the unemployment rate of KLF against LALF for urban areas. From Figure 5.1, we can find that there is a linear structural relationship between KLF and LALF. Instead of the usual residual $\hat{\bar{e}}_h$ in the structural error model, $\hat{v}_h$ are used as the residuals in the regression model with measurement errors, where $\hat{v}_h = \bar{y}_{1h} - \hat{\beta}_1 \bar{x}_h$. Figure 5.2 contains a plot of $\hat{v}_h$ against $\hat{\bar{X}}_h$ for urban area. The plot shows that the assumption of equal variance σ_e^2 is slightly violated. The heteroscedastic variance model in Remark 2 was also considered but the results did not change significantly.

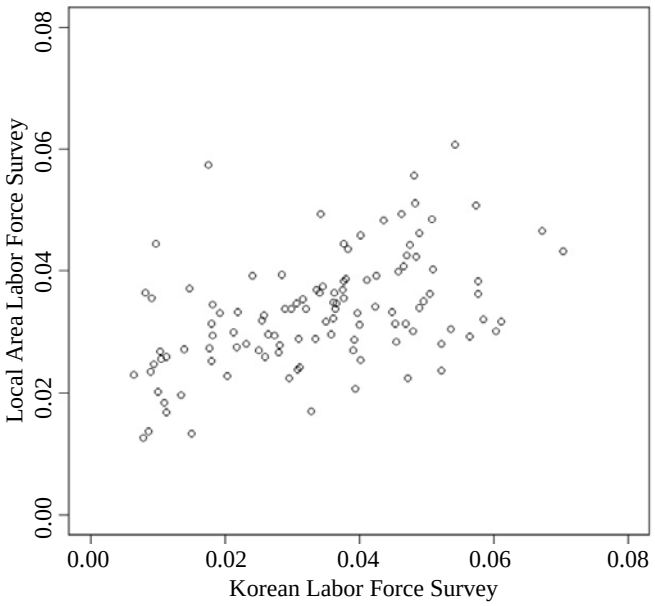

Figure 5.1 Plot of unemployment rate for KLF and LALF survey for urban area.

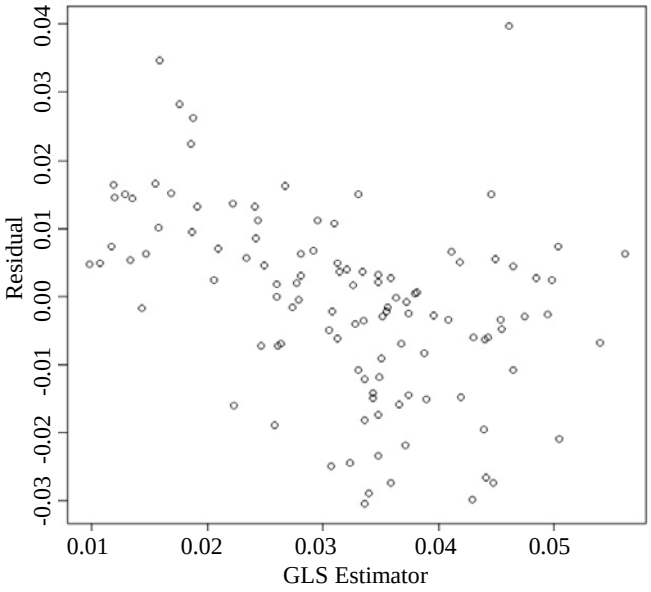

Figure 5.2 Plot of residuals against estimated values for urban area.

Table 5.1 presents the performance of the small area estimates in terms of the MSE estimates. We considered four different estimators of $\bar{X}_h$. KLF represents the result derived using only Korea Labor Force survey, LALF represents the result using only Local Area Labor Force survey, GLS 1 represents the

result for combining both surveys KLF and LALF, and GLS 2 represents the result for combining KLF, LALF and the Census data. Table 5.1 shows that the GLS 2 method provides the smallest mean squared errors.

Table 5.1
Quartile of the MSE performance of the small area estimates for the 229 areas

MSE	1st Q	Median	3rd Q	Mean
KLF	0.0000630	0.0001210	0.0002395	0.0002476
LALF	0.0001123	0.0001330	0.0001695	0.0001482
GLS 1	0.0000444	0.0000738	0.0001210	0.0000893
GLS 2	0.0000405	0.0000543	0.0000721	0.0000575

6 Concluding remark

In this paper, a small area estimation problem is treated as a measurement error model prediction problem where the covariates, which are the direct estimates for small areas, are subject to sampling errors. In our measurement error model approach, the sampling errors of the direct estimators are treated as measurement errors and the structural error model can be used to link the other auxiliary estimates to the direct estimators. The proposed model is actually the opposite of the model of Ybarra and Lohr (2008), where the direct estimator is treated as a dependent variable in the regression model and the nonsampling errors of auxiliary estimates are treated as measurement errors.

In our approach, each auxiliary estimate is treated as a dependent variable in the regression model using the direct estimate as the covariate and the sampling error of the direct estimator is treated as measurement error. The measurement error variance is easy to estimate because it is essentially the sampling variance of the direct estimate. The measurement error model approach is also very useful when there are several sources of auxiliary information of area-levels. Unlike the Bayesian approach, the resulting estimator does not rely on parametric model assumptions about the structural error model and is still optimal in the sense of minimizing the mean squared errors among the class of unbiased estimators that are linear in the available data.

In the example of the Korean labor survey application, two sample estimates and the Census information are used to compute the GLS estimates for small area parameters and the two sample estimates are correlated due to the two-phase sampling structure. We simply used linear regression models for the linking models, mainly for the sake of computational simplicity. Instead of the linear model, one may consider a generalized linear model to improve model prediction power. Such extension would involve the theory for nonlinear measurement error models. Further investigation on this extension will be a topic of future research.

Acknowledgements

We thank an anonymous referee and the Associate Editor for their constructive comments. The research of the first author was partially supported by a grant from NSF (MMS-121339).

Appendix

Reversed two-phase sampling

In the classical two-phase sampling, the second-phase sample (A_2) is a subset of the first-phase sample (A_1). We consider another type of sampling design that has a reversed structure of the two-phase sampling design. In the reversed two-phase sampling design, we have the following sampling steps:

Step 1 From the finite population, we select the first-phase sample A_1 of size n_1.

Step 2 In the second-phase sample, we select A_2 from $U - A_1$ of size n_2. The final sample A consists of A_1 and A_2. That is, $A = A_1 \cup A_2$ and $|A| = n = n_1 + n_2$.

The reversed two-phase sampling is used when the sample is augmented by an additional sampling procedure.

To discuss parameter estimation under reversed two-phase sampling, let $\pi_{1i} = \Pr(i \in A_1)$ be the first-order inclusion probability for A_1. Let $\pi_{2i|1} = \Pr\left(i \in A_2 | A_1^c\right)$ be the conditional first-order inclusion probability for A_2 given $A_1^c = U - A_1$. To compute the inclusion probability for A,

$$\Pr(i \in A) = \Pr(i \in A_1) + \Pr\left(i \in A_2 | A_1^c\right) \Pr\left(i \in A_1^c\right).$$

Thus, we can use $\pi_i = \pi_{1i} + (1 - \pi_{1i})\pi_{2i|1}$ to compute the Horvitz-Thompson estimator of the form

$$\hat{Y}_{r,\mathrm{HT}} = \sum_{i \in A} \frac{1}{\pi_i} y_i. \tag{A.1}$$

Note that, instead of (A.1), we can consider the following class of estimators:

$$\hat{Y}_w = W \sum_{i \in A_1} \frac{1}{\pi_{1i}} y_i + (1 - W) \sum_{i \in A_2} \frac{1}{\pi_{2i|1}(1 - \pi_{1i})} y_i := W\hat{Y}_1 + (1 - W)\hat{Y}_2. \tag{A.2}$$

Since $\hat{Y}_1$ and $\hat{Y}_2$ are both unbiased for Y, $\hat{Y}_w$ is also unbiased regardless of the choice of W. A reasonable choice of W is $W = n_1/n$.

Under simple random sampling in both designs, the two estimators are equal to $\hat{Y} = N\bar{y}_n$, where $\bar{y}_n$ is the sample mean of y in A. Writing $\bar{y}_1 = n_1^{-1} \sum_{i \in A_1} y_i$ and $\bar{y}_2 = \sum_{i \in A_2} y_i / n_2$, we have

$$\bar{y}_n = W\bar{y}_1 + (1 - W)\bar{y}_2 \tag{A.3}$$

where $W = n_1/n$. Using

$$V(\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 \tag{A.4}$$

$$V(\bar{y}_2) = \left(\frac{1}{n_2} - \frac{1}{N}\right) S_y^2$$

$$\text{Cov}(\bar{y}_1, \bar{y}_2) = \text{Cov}\left(\bar{y}_1, \bar{y}_1^c\right) = -\frac{n_1}{N - n_1}\left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 = -\frac{1}{N} S_y^2,$$

where $\bar{y}_1^c = \sum_{i \in A_1^c} y_i / (N - n_1)$, we have, for $W = n_1 / n$,

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.5}$$

Also,

$$\text{Cov}(\bar{y}_1, \bar{y}_n) = \text{Cov}[\bar{y}_1, W\bar{y}_1 + (1 - W)\bar{y}_2] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.6}$$

If $W = n_1 / n$ does not hold, then (A.5) and (A.6) do not hold.

In the KLF application in Section 5, since x and y are measuring the same item, we may assume $S_x^2 = S_y^2 = S_{xy}$ and the variance-covariance matrix of the sampling errors can be smoothed as

$$V(a_h, b_h) = \begin{pmatrix} n_1^{-1} & n^{-1} \\ n^{-1} & n^{-1} \end{pmatrix} S_y^2.$$

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.

Carroll, R.J., Rupert, D. and Stefanski, L.A. (1995). *Measurement error in nonlinear models*. New York: Chapman & Hall.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fuller, W.A. (1987). *Measurement error models*. New York: John Wiley & Sons, Inc.

Fuller, W.A. (1991). Small area estimation as a measurement error problem. In *Economic Models, Estimation, and Socioeconomic Systems: Essays in Honor of Karl A. Fox*, (Eds., Tij K. Kaul and Jati K. Sengupta), Elsevier Science Publishers, 333-352.

Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, NJ.

Jiang, J., Lahiri, P. and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics*, 30, 1782-1810.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association,* 79, 853-862.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 85-100.

Lohr, S.L., and Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics,* 31, 383-396.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. and Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A,* 174, 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society B,* 68, 509-521.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review,* 70, 125-144.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika,* 43, 353-360.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.I., Davis, W.W., Dodd, K.W. and Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association,* 102, 474-486.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, NJ.

Schafer, D.W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics,* 57, 53-61.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika,* 95, 919-931.