CONFIDENCE INTERVALS FOR UNIT AND AREA LEVEL SMALL AREA ESTIMATORS

Mike Hidiroglou and Yong You[1]

ABSTRACT

In this paper, we summarize the EBLUP and pseudo-EBLUP estimators for small area estimation under a basic unit level model and the EBLUP estimators under the Fay-Herriot area level model. In particular, we are interested in the confidence interval coverage of the EBLUP and pseudo-EBLUP estimates. We conducted a design-based simulation study to compare the model-based estimates for unit level and area level models. We also compared the estimates under certain model misspecification. Our results have shown that unit level model performs better than the area level model under correct modeling, and pseudo-EBLUP estimator is the best among unit level and area level estimators.

KEY WORDS: Benchmarking, Design consistent, Nested error regression model, MSE, Survey weight.

RÉSUMÉ

Dans cet article, nous résumons les estimateurs « EBLUP » et « pseudo- EBLUP « pour l'estimation des petits domaines en supposant un modèle au niveau de l'unité et les estimateurs EBLUP dans le cadre du modèle de Fay-Herriot. En particulier, nous nous intéressons à la couverture de l'intervalle de confiance des estimations du l'EBLUP ainsi que du pseudo- EBLUP. Nous avons mené une étude de simulation basé sur un plan de sondage afin comparer les estimations fondées sur un modèle au niveau de l'unité ainsi qu'au au niveau de la région. Nous avons également comparé les estimations selon de mauvaises spécifications du modèle. Nos résultats ont démontré que le modèle au niveau de l'unité est plus performant que celui au niveau de la région en supposant que le modèle est juste. Aussi, l'estimateur pseudo- EBLUP est le meilleur parmi ceux au niveau de l'unité ainsi que ceux au niveau de la de la région.

MOTS CLÉS: Calage; consistent au niveau du plan de sondage modèle de régression emboitée, EQM, Poids de sondage)

1. INTRODUCTION

Small area estimation is carried out using models that can be classified into two broad types: (i) Aggregate level (or area level) models that relate the small area means to area-specific auxiliary variables. ii) Unit (element) level models that relate the unit values of the study variable to unit-specific auxiliary variables. Area level models will be used if unit level data are not available. The common model that drives these procedures is the General Linear Mixed Model (GLMM) given by $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + e$, where **y** is the n x 1 vector of sample observations, X and Z are known n x p and n x h matrices of full rank, and v and e are independently distributed.

The survey design can be incorporated into these broad types in different ways. In the case of area level, the survey variance of the associated direct estimator is introduced into the model via the design-induced errors **e** in the GLMM. In the case of the unit level, the observations can be weighted with the survey weight. A number of factors affect the success of using these estimators. Two important factors are whether the assumed model is correct and whether the variable of interest is correlated with the selection probabilities associated with the sampling process (informativeness). In this paper, we compare, via simulation, the impact of model misspecification and informativeness for these two basic procedures in terms of bias, estimated mean squared error and coverage.

The paper is structured as follows. The point and associated estimated mean squared errors for the unit level and area models are described in sections 2 and 3 respectively. The description of the simulation and results are given in section 4. This simulation computes the point and associated mean squared errors for a ppswr sampling scheme for the following two conditions: a. the model is correct or incorrect; and b. Design informativeness varies from insignificant to very significant. Finally, conclusions resulting from this work are discussed in Section 5.

[1] Statistical Research and Innovation Division, Statistics Canada, Ottawa, K1A 0T6, Canada.

2. UNIT LEVEL MODEL

A basic unit level model for small area estimation is the nested error regression model (Battese, Harter and Fuller, 1988) give as $y_{ij} = \boldsymbol{x}_{ij}'\beta + v_i + e_{ij}, \quad j = 1,...,N_i, i = 1,...,m,$ where y_{ij} is the variable of interest for the j-th population unit in the i-th small area, $\boldsymbol{x}_{ij} = (x_{ij1},...,x_{ijp})'$ with $x_{ij1} = 1$ is a $p \times 1$ vector of auxiliary variables associated with y_{ij}, $\boldsymbol{\beta} = (\beta_0,...,\beta_{p-1})'$ is a $p \times 1$ vector of regression parameters, and N_i is the number of population units in the i-th small area. The random effects v_i are assumed to be $iid\ N(0,\sigma_v^2)$ and independent of the unit errors e_{ij}, which are assumed to be $iid\ N(0,\sigma_e^2)$. The parameter of interest is the mean for the i-th area, $\bar{Y}_i$, which may be approximated by

$$\theta_i = \bar{\boldsymbol{X}}_i'\boldsymbol{\beta} + v_i, \tag{1}$$

assuming that N_i is large, where $\bar{\boldsymbol{X}}_i'$ is the vector of known population means of the $\boldsymbol{x}_{ij}$ for the i-th area, that is, $\bar{\boldsymbol{X}}_i = \sum_{j=1}^{N_i} \boldsymbol{x}_{ij} / N_i$. We assume that samples are drawn independently across small areas according to a specified sampling design. The sample data $\{y_{ij}, x_{ij}, \tilde{w}_{ij}, j = 1,...,n_i; i = 1,...,m\}$ is assumed to obey the population model, i.e.

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1,...,n_i, i = 1,...,m, \tag{2}$$

where $\tilde{w}_{ij}$ is the basic design weight associated with unit y_{ij}, and n_i is the sample size in the i-th small area.

2.1 EBLUP Estimation

The best linear unbiased prediction (BLUP) estimator of small area mean $\theta_i = \bar{\boldsymbol{X}}_{\boldsymbol{i}}'\boldsymbol{\beta} + v_i$ based on the nested error regression model (2) is given by

$$\tilde{\theta}_i = r_i \bar{y}_i + (\bar{\boldsymbol{X}}_i - r_i\, \bar{\boldsymbol{x}}_i)'\tilde{\boldsymbol{\beta}}, \tag{3}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$, $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} / n_i$, $r_i = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 / n_i)$, and $\tilde{\boldsymbol{\beta}}$ is given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \bar{\boldsymbol{x}}_i' \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \bar{\boldsymbol{x}}_i' \boldsymbol{V}_i^{-1} y_i \right) \equiv \tilde{\boldsymbol{\beta}}(\sigma_e^2, \sigma_v^2), \tag{4}$$

where $\boldsymbol{x}_i' = (x_{i1},...,x_{in_i})$, $\boldsymbol{V}_i = \sigma_e^2 \boldsymbol{I}_{n_i} + \sigma_v^2 \boldsymbol{1}_{n_i} \boldsymbol{1}_{n_i}'$, $y_i = (y_{i1},...,y_{in_i})'$, $i = 1,...,m$. Both $\tilde{\theta}_i$ and $\tilde{\boldsymbol{\beta}}$ depend on the unknown variance parameters σ_e^2 and σ_v^2. We use the method of fitting constant to estimate σ_e^2 and σ_v^2, and the estimators are given as

$$\hat{\sigma}_e^2 = (n - m - p + 1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2,$$

and $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$, where $\tilde{\sigma}_v^2$ is given by

$$\tilde{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n - p)\hat{\sigma}_e^2 \right],$$

where $n_* = n - tr[(\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{i=1}^{m} n_i^2 \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i']$, $\boldsymbol{X}' = (\boldsymbol{x}_1',...,\boldsymbol{x}_m')$. Furthermore, $\{\hat{\varepsilon}_{ij}\}$ are residuals from the ordinary least squares (OLS) regression of $y_{ij} - \bar{y}_i$ on $\{\boldsymbol{x}_{ij1} - \bar{\boldsymbol{x}}_{i\cdot 1},..., \boldsymbol{x}_{ijp} - \bar{\boldsymbol{x}}_{i\cdot p}\}$ and $\{\hat{u}_{ij}\}$ are the residuals from the OLS regression of y_{ij} on $\{x_{ij1},..., x_{ijp}\}$. See Rao (2003), page 138 for more details.

Replacing σ_e^2 and σ_v^2 by estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, we obtain the EBLUP estimator of small area mean θ_i as

$$\hat{\theta}_i = r_i \bar{y}_i + (\bar{\boldsymbol{X}}_i - \hat{r}_i\, \bar{\boldsymbol{x}}_i)'\hat{\boldsymbol{\beta}} \quad , \tag{5}$$

where $\hat{r}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$. The mean squared error (MSE) of the EBLUP estimator $\hat{\theta}_i$ is given as $MSE(\hat{\theta}_i) \approx g_{1i}(\sigma_e^2, \sigma_v^2) + g_{2i}(\sigma_e^2, \sigma_v^2) + g_{3i}(\sigma_e^2, \sigma_v^2)$, see Prasad and Rao (1990). The g-terms are $g_{1i}(\sigma_e^2, \sigma_v^2) = (1 - r_i)\sigma_v^2$,

$g_{2i}(\sigma_e^2,\sigma_v^2) = (\bar{\boldsymbol{X}}_i - r_i\,\bar{\boldsymbol{x}}_i)'(\sum_{i=1}^m \boldsymbol{x}_i'\boldsymbol{V}_i^{-1}\boldsymbol{x}_i)^{-1}(\bar{\boldsymbol{X}}_i - r_i\,\bar{\boldsymbol{x}}_i)$ and $g_{3i}(\sigma_e^2,\sigma_v^2) = n_i^{-2}(\sigma_v^2+\sigma_e^2 n_i^{-1})^{-3}h(\sigma_e^2,\sigma_v^2)$. Here, $h(\sigma_e^2,\sigma_v^2)$ is $h(\sigma_e^2,\sigma_v^2) = \sigma_e^4 V(\tilde{\sigma}_v^2) - 2\sigma_e^2\sigma_v^2\,\mathrm{cov}(\hat{\sigma}_e^2,\tilde{\sigma}_v^2) + \sigma_v^4 V(\hat{\sigma}_e^2)$. The variances and covariance of $\hat{\sigma}_e^2$ and $\tilde{\sigma}_v^2$ are given as following terms: $V(\hat{\sigma}_e^2) = 2(n-m-p+1)^{-1}\sigma_e^4$, $V(\tilde{\sigma}_v^2) = 2n_*^{-2}[(n-m-p+1)^{-1}(m-1)(n-p)\sigma_e^4 + 2n_*\sigma_e^2\sigma_v^2 + n_{**}\sigma_v^4]$, and $\mathrm{cov}(\hat{\sigma}_e^2,\tilde{\sigma}_v^2) = -(m-1)n_*^{-1}V(\hat{\sigma}_e^2)$, where $n_* = n - tr[(X'X)^{-1}\sum_{i=1}^m n_i^2\,\bar{\boldsymbol{x}}_i\,\bar{\boldsymbol{x}}_i']$, $n_{**} = tr(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^2$, $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{Z} = diag(\boldsymbol{1}_{n_1},...,\boldsymbol{1}_{n_m})$. A second-order unbiased estimator of the MSE is obtained by Prasad and Rao (1990) as

$$mse(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_e^2,\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_e^2,\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_e^2,\hat{\sigma}_v^2). \tag{6}$$

Note that the EBLUP estimator $\hat{\theta}_i$ given by (5) depends on the unit level model (2) is generally not design consistent. If the model (2) does not hold for the sampled data, then the EBLUP estimator $\hat{\theta}_i$ may lead to biased estimates.

2.2 Pseudo-EBLUP Estimation

You and Rao (2002) proposed a pseudo-EBLUP estimator of the small area mean θ_i by combining the survey weights and the unit level model (2) to achieve design consistency. Let $\tilde{w}_{ij}$ be the weights associated with each observation unit y_{ij}. A direct design-based estimator of the small area mean is given by

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i}\tilde{w}_{ij}y_{ij}}{\sum_{j=1}^{n_i}\tilde{w}_{ij}} = \sum_{j=1}^{n_i} w_{ij}y_{ij}, \tag{7}$$

where $w_{ij} = \tilde{w}_{ij}/\sum_{j=1}^{n_i}\tilde{w}_{ij} = \tilde{w}_{ij}/\tilde{w}_{i\cdot}$ and $\sum_{j=1}^{n_i} w_{ij} = 1$. By combining the direct estimator (7) and the unit level model (2), we can obtain the following aggregated (survey-weighted) area level model:

$$\bar{y}_{iw} = \bar{\boldsymbol{x}}_{iw}'\boldsymbol{\beta} + v_i + \bar{e}_{iw}, \quad i = 1,...,m, \tag{8}$$

where $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij}e_{ij}$ with $E(\bar{e}_{iw}) = 0$ and $V(\bar{e}_{iw}) = \sigma_e^2\sum_{j=1}^{n_i} w_{ij}^2 \equiv \delta_i^2$, and $\bar{\boldsymbol{x}}_{iw} = \sum_{j=1}^{n_i} w_{ij}\boldsymbol{x}_{ij}$. Note that the regression parameter $\boldsymbol{\beta}$ and the variance components σ_e^2 and σ_v^2 are unknown in model (8). Based on model (8), assuming that the parameters $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 are known, we can obtain the BLUP estimator of θ_i as

$$\tilde{\theta}_{iw} = r_{iw}\bar{y}_{iw} + (\bar{\boldsymbol{X}}_i - r_{iw}\,\bar{\boldsymbol{x}}_{iw})'\boldsymbol{\beta} = \tilde{\theta}_{iw}(\boldsymbol{\beta},\sigma_e^2,\sigma_v^2), \tag{9}$$

where $r_{iw} = \sigma_v^2/(\sigma_v^2+\delta_i^2)$. The BLUP estimator $\tilde{\theta}_{iw}$ depends on $\boldsymbol{\beta}$, σ_e^2 and σ_v^2. To estimate the regression parameter, You and Rao (2002) proposed a weighted estimation equation approach, and obtained an estimator as follows:

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{w}_{ij}\boldsymbol{x}_{ij}(\boldsymbol{x}_{ij} - r_{iw}\,\bar{\boldsymbol{x}}_{iw})'\right]^{-1}\left[\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{w}_{ij}(\boldsymbol{x}_{ij} - r_{iw}\,\bar{\boldsymbol{x}}_{iw})y_{ij}\right] \equiv \tilde{\boldsymbol{\beta}}_w(\sigma_e^2,\sigma_v^2). \tag{10}$$

Note that $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\sigma_e^2,\sigma_v^2)$ depends on σ_e^2 and σ_v^2. Replacing σ_e^2 and σ_v^2 in (10) by the fitting of constant estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, we can obtain $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\hat{\sigma}_e^2,\hat{\sigma}_v^2)$. See Rao (2003, page 149).

Now replacing $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 in (9) by $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, we obtain the pseudo-EBLUP estimator for the small area mean θ_i:

$$\hat{\theta}_{iw} = \hat{r}_{iw}\bar{y}_{iw} + (\bar{\boldsymbol{X}}_i - \hat{r}_{iw}\,\bar{\boldsymbol{x}}_{iw})'\hat{\boldsymbol{\beta}}_w.$$

Note that $\hat{\theta}_{iw}$ is design-consistent as the sample size n_i becomes large. Also, $\hat{\theta}_{iw}$ has a nice self-benchmarking property assuming that the weights $\tilde{w}_{ij}$ are calibrated to agree with the known population total, that is, if $\sum_{j=1}^{n_i}\tilde{w}_{ij} = N_i$, then $\sum_{i=1}^m N_i\hat{\theta}_{iw}$ equals the direct regression estimator of the overall total, that is,

$$\sum_{i=1}^m N_i\hat{\theta}_{iw} = \hat{Y}_w + (\boldsymbol{X} - \hat{\boldsymbol{X}}_w)'\hat{\boldsymbol{\beta}}_w,$$

where $\hat{Y}_w = \sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{w}_{ij}y_{ij}$, and $\hat{\boldsymbol{X}}_w = \sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{w}_{ij}\boldsymbol{x}_{ij}$. For more details, see You and Rao (2002).

The MSE of $\hat{\theta}_{iw}$ is given as

$$MSE(\hat{\theta}_{iw}) \approx g_{1iw}(\sigma_e^2, \sigma_v^2) + g_{2iw}(\sigma_e^2, \sigma_v^2) + g_{3iw}(\sigma_e^2, \sigma_v^2),$$

where $g_{1iw}(\sigma_e^2, \sigma_v^2) = (1 - r_{iw})\sigma_v^2$, $g_{2iw}(\sigma_e^2, \sigma_v^2) = (\bar{X}_i - r_{iw}\bar{x}_{iw})'\Phi_w(\bar{X}_i - r_{iw}\bar{x}_{iw})$, and Φ_w is given as

$$\Phi_w(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}z_{ij}')^{-1}(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}')[(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}')^{-1}]'\sigma_e^2 + (\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}')^{-1}[\sum_{i=1}^{m}(\sum_{j=1}^{n_i}\boldsymbol{z}_{ij})(\sum_{j=1}^{n_i}\boldsymbol{z}_{ij})'][(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}\boldsymbol{z}_{ij})^{-1}]'\sigma_v^2,$$

and $\boldsymbol{z}_{ij} = \tilde{w}_{ij}(\boldsymbol{x}_{ij} - r_{iw}\bar{\boldsymbol{x}}_{iw})$, $g_{3iw}(\sigma_e^2, \sigma_v^2) = r_{iw}(1 - r_{iw})^2\sigma_e^{-4}\sigma_v^{-2}h(\sigma_e^2, \sigma_v^2)$, where $h(\sigma_e^2, \sigma_v^2)$ is the same function as in the MSE for the EBLUP estimator. A nearly second-order unbiased estimator of the MSE can be obtained as

$$mse(\hat{\theta}_{iw}) = g_{1iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + g_{2iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + 2g_{3iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2). \quad (11)$$

See Rao (2003, page 150) and You and Rao (2002, page 435). Note that this MSE estimator (11) ignores the cross-product terms. However, the cross-product terms are relatively small; see Torabi and Rao (2010) for details.

Note that the pseudo-EBLUP estimator $\hat{\theta}_{iw}$ is slightly less efficient than the EBLUP estimator $\hat{\theta}_i$, but the pseudo-EBLUP estimator is design consistent and will be more robust against model misspecification. We will compare the performance of the EBLUP and pseudo-EBLUP estimators through a simulation study.

3. AREA LEVEL MODEL

The Fay-Herriot model (Fay and Herriot, 1979) is a basic area level model widely used in small area estimation to improve the direct survey estimates. The Fay-Herriot model basically has two components, namely, a sampling model for the direct survey estimates and a linking model for the small area parameters of interest. The sampling model assumes that given the area-specific sample size $n_i > 1$, there exists a direct survey estimator y_i, which is usually design unbiased, for the small area parameter θ_i such that

$$y_i = \theta_i + e_i,\ i = 1, ..., m, \quad (12)$$

where the e_i is the sampling error associated with the direct estimator y_i and m is the number of small areas. It is customary in practice to assume that the e_i's are independently normal random variables with mean $E(e_i|\theta_i) = 0$ and sampling variance $\text{var}(e_i|\theta_i) = \sigma_i^2$. To obtain the linking model we assume that the small area parameter of interest θ_i is related to area level auxiliary variables $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})'$ through a linear regression model

$$\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i,\ i = 1, ..., m, \quad (13)$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a p × 1 vector of regression coefficients, and the v_i's are area-specific random effects assumed to be independent and identically distributed (iid) with E(v_i) = 0 and var(v_i) = σ_v^2. The assumption of normality is generally also included, even though it is more difficult to justify the assumption. The model variance σ_v^2 is unknown and needs to be estimated from the data. The area level random effects v_i capture the unstructured heterogeneity among areas that are not explained by the sampling variances. Combining models (12) and (13) lead to a linear mixed area level model given as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + e_i, \quad (14)$$

Model (14) involves both design-based random errors e_i and model-based random effects v_i. For the Fay-Herriot model, the sampling variance σ_i^2 is assumed to be known in model (14). This is a very strong assumption. Generally smoothed estimators of the sampling variances are used in the Fay-Herriot model and then treated as known. However, if direct estimates of sampling variances are used in the Fay-Herriot, then an extra term is added to the MSE estimator to account for the extra variation (Wang and Fuller, 2003).

Assuming the model variance σ_v^2 is known, the best linear unbiased predictor (BLUP) of the small area parameter θ_i can be obtained as

$$\tilde{\theta}_i = \gamma_i y_i + (1-\gamma_i)\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}_{WLS}, \quad (15)$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2+\sigma_i^2)$, and $\tilde{\boldsymbol{\beta}}_{WLS}$ is the weighted least squared (WLS) estimator of $\boldsymbol{\beta}$ given as

$$\tilde{\boldsymbol{\beta}}_{WLS} = \left[\sum_{i=1}^{m}(\sigma_i^2+\sigma_v^2)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{m}(\sigma_i^2+\sigma_v^2)^{-1}\boldsymbol{x}_i y_i\right] = \left[\sum_{i=1}^{m}\gamma_i\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{m}\gamma_i\boldsymbol{x}_i y_i\right].$$

To estimate the unknown model variance σ_v^2, there are several methods available; see You (2010) for a review. We consider the restricted maximum likelihood (REML) method. The REML method is derived by Cressie (1992) to estimate the model variance under the Fay-Herriot model. By using the scoring algorithm, the REML estimator $\hat{\sigma}_v^2$ is obtained as follows (Cressie, 1992; Rao, 2003):

$$\sigma_v^{2(k+1)} = \sigma_v^{2(k)} + \left[I_R(\sigma_v^{2(k)})\right]^{-1} S_R(\sigma_v^{2(k)}), \text{ for } k = 1,2,\ldots,$$

where $I_R(\sigma_v^2) = \frac{1}{2}\mathrm{tr}[\mathbf{PP}]$, and $S_R(\sigma_v^2) = \frac{1}{2}\mathbf{y}'\mathbf{PPy} - \frac{1}{2}\mathrm{tr}[\mathbf{P}]$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. Note that we can simply use a guess value as the starting value for $\sigma_v^{2(1)}$. The algorithm should converge very fast.

Replacing σ_v^2 in equation (15) by the REML estimator $\hat{\sigma}_v^2$, we can obtain the EBLUP of the small area parameter θ_i based on the Fay-Herriot model as

$$\hat{\theta}_i^{FH} = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{WLS},$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2+\sigma_i^2)$. As in the unit level model, and from Prasad and Rao (1990) and Rao (2003), we can obtain the MSE estimator of $\hat{\theta}_i^{FH}$ as follows:

$$mse(\hat{\theta}_i^{FH}) = g_{1i} + g_{2i} + 2g_{3i},$$

where g_{1i} is the leading term given as $g_{1i} = \hat{\gamma}_i\sigma_i^2$, g_{2i} accounts for the variability due to estimation of the regression parameter β, and is given by

$$g_{2i} = (1-\hat{\gamma}_i)^2\boldsymbol{x}_i' Var(\hat{\boldsymbol{\beta}}_{WLS})\boldsymbol{x}_i = \hat{\sigma}_v^2(1-\hat{\gamma}_i)^2\boldsymbol{x}_i'\left(\sum_{i=1}^{m}\hat{\gamma}_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\boldsymbol{x}_i.$$

The term g_{3i} is due to the estimation of the model variance and given as $g_{3i} = (\sigma_i^2)^2(\hat{\sigma}_v^2+\sigma_i^2)^{-3}V(\hat{\sigma}_v^2)$, where $V(\hat{\sigma}_v^2)$ is the asymptotic variance of the REML estimator $\hat{\sigma}_v^2$ obtained by Datta and Lahiri (2000) given by $V(\hat{\sigma}_v^2) = 2(\sum_{i=1}^{m}(\hat{\sigma}_v^2+\sigma_i^2)^{-2})^{-1}$.

In the above discussions, the sampling variance σ_i^2 is assumed to be known in the Fay-Herriot model (14). This is very strong assumption. Usually a direct survey estimator, say s_i^2, of the sampling variance σ_i^2 is available. The direct sampling variance estimates are smoothed as $\tilde{s}_i^2$ by using external models and generalized variance functions. The smoothed sampling variance estimates $\tilde{s}_i^2$ are used in the Fay-Herriot model and treated as known. Rivest and Vandal (2003) and Wang and Fuller (2003) considered the small area estimation using the Fay-Herriot model with the direct sampling variance estimates s_i^2 under the assumption that the estimators s_i^2 are independent of the direct survey estimators y_i and $d_i s_i^2 \sim \sigma_i^2\chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the *i*-th area. When the direct sampling variance estimate s_i^2 is used in the place of the true sampling variance σ_i^2, an extra term accounts for the uncertainty of using s_i^2 is needed in the MSE estimator, and this term, denoted as g_{4i}, is given as

$$g_{4i} = \frac{4}{n_i - 1}\frac{\hat{\sigma}_v^4 s_i^4}{(\hat{\sigma}_v^2+s_i^2)^3};$$

see Rivest and Vandal (2003) and Wang and Fuller (2003) for details.

We will conduct a simulation study to compare the Fay-Herriot model with the unit level model, particularly the bias and confidence interval of the model-based estimates.

4. SIMULATION STUDY

4.1 Data Generation

We created two finite populations. Each finite population had $m = 30$ areas, and each area consisted of $N_i = 200$ population units. The first finite population was generated from the unit level model $y_{ij} = \beta_0 + x_{1ij}\beta_1 + v_i + e_{ij}$ by taking $\beta_0 = 50$, $\beta_1 = 10$, $v_i \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, where $\sigma_e^2 = 225$ and $\sigma_v^2 = 100$. The auxiliary variable x_{1ij} was generated from an exponential distribution with mean 4 and variance 8. The second population was generated from the same model but with different fixed effects values: $\beta_0 = 50$, $\beta_1 = 10$ for areas $m = 1, ..., 10$; $\beta_0 = 75$, $\beta_1 = 15$ for areas $m = 11, ..., 20$; $\beta_0 = 100$, $\beta_1 = 20$ for areas $m = 21, ..., 30$. Therefore, in the second population, we had three different means for the fixed effects $\beta_0 + x_{1ij}\beta_1$. From the constructed populations, PPSWR (probability proportional to size with replacement) samples within each area were drawn independently. To implement PPSWR sampling, we defined a size measure z_{ij} for each y_{ij}. Using these z_{ij} values, we computed selection probabilities $p_{ij} = z_{ij} / \sum_j z_{ij}$ for each unit y_{ij} and used them to select PPSWR samples of equal size, $n_i = n$, within each group, by taking n = 10 and 30, respectively. The basic design weights are given by $\tilde{w}_{ij} = n_i^{-1} p_{ij}^{-1}$ so that $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$. We chose the size measure z_{ij} in a way such that the correlation coefficient between y_{ij} and selection probability p_{ij} within each group varied between 0.01 to 0.95, that corresponds to non-informative selection to strongly informative selection of PPSWR samples.

4.2 Model Fitting

For unit level modeling, we fitted the nested error regression model to the PPSWR sampling data generated from the population. We then obtained the corresponding EBLUP and pseudo-EBLUP estimates and related confidence interval coverage estimates. For the area level Fay-Herriot model, we first constructed three area level direct estimates and the corresponding sampling variance estimates using the selected PPS samples. Table 1 presents these estimates. Then we used these area level estimates as input values and fit the area level Fay-Herriot model, and obtained three area level model-based estimates correspondingly denoted as: FH-SRS, FH-HT, and FH-HA. Note that the SRS direct estimator ignored the sample design and should perform poorly if the sample design is informative. The weighted Hájek estimator is also used in the pseudo-EBLUP estimator for the unit level model.

Table 1: Areal level direct estimator and sampling variances

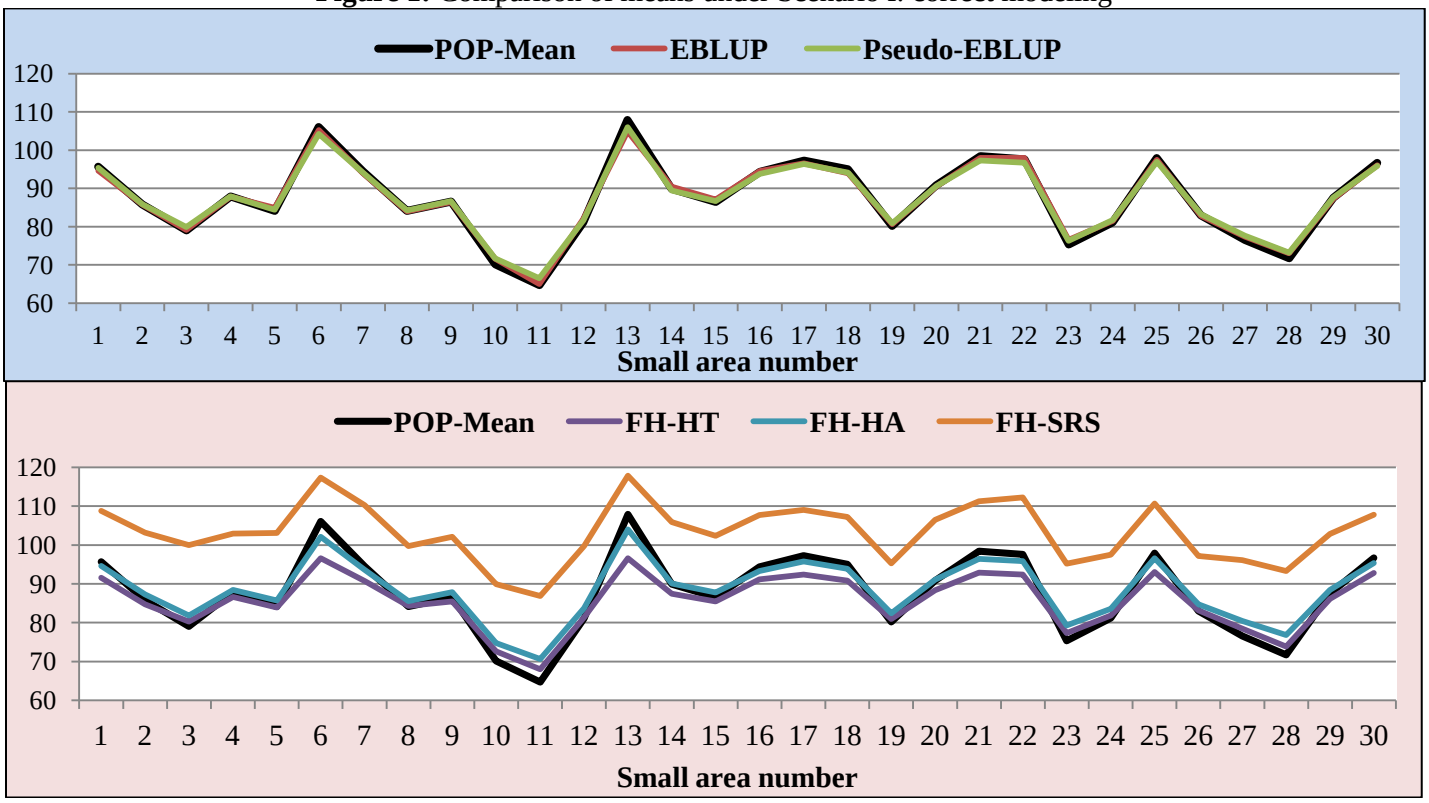
	Point estimator	Sampling variance estimator
Direct mean (SRS)	$\hat{\theta}_i^{SRS} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	$\text{var}(\hat{\theta}_i^{SRS}) = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{\theta}_i^{SRS}\right)^2$
Horvitz-Thompson (HT) estimator	$\hat{\theta}_i^{HT} = \frac{1}{N_i}\sum_{j=1}^{n_i}\tilde{w}_{ij} y_{ij} = \frac{1}{N_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}(\hat{\theta}_i^{HT}) = \frac{1}{N_i^2 n_i (n_i - 1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij}}{p_{ij}} - N_i\,\hat{\theta}_i^{HT}\right)^2$
Weighted Hájek (HA) estimator	$\hat{\theta}_i^{HA} = \frac{\sum_{j=1}^{n_i}\tilde{w}_{ij} y_{ij}}{\sum_{j=1}^{n_i}\tilde{w}_{ij}} = \frac{1}{\hat{N}_i}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i p_{ij}}$	$\text{var}(\hat{\theta}_i^{HA}) = \frac{1}{\hat{N}_i^2 n_i (n_i - 1)}\sum_{j=1}^{n_i}\left(\frac{y_{ij} - \hat{\theta}_i^{HA}}{p_{ij}}\right)^2$

For both unit level and area level model fitting, we had the following two scenarios: Scenario (I): correct modeling, where the data is generated from the first population and the fitting models were unit level model (2) and area level model (14) with common β_0 and β_1. Scenario (II): incorrect modeling, where the data is generated from the second population with different means for the fixed effects, and the fitting models were the same as in Scenario (I) with common β_0 and β_1.

4.3 Results

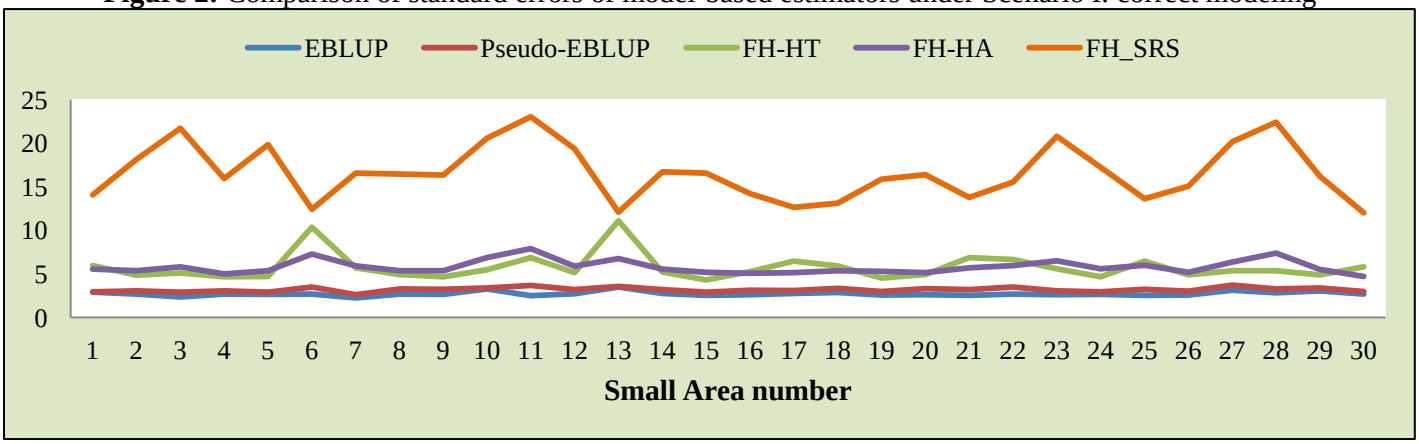
We first compared the means and standard errors of the unit level and area level estimates under Scenario I: correct modeling. Figure 1 presents the comparison of unit and area level estimates with the population means when the PPSWR sampling procedure is informative (correlation coefficient between y_{ij} and the selection probability p_{ij} is 0.8).

Figure 1: Comparison of means under Scenario I: correct modeling



It is clear that for correct modeling, both EBLUP and pseudo-EBLUP lead to unbiased estimates. FH-SRS severely overestimates the means, and both FH-HT and FH-HA lead to reasonable estimates, but FH-HT has larger bias than FH-HA. Figure 2 presents the comparison of standard errors for both unit and area level estimators under correct modeling. EBLUP and pseudo-EBLUP have much smaller standard errors than the FH area level estimates. EBLUP has the smallest standard errors and pseudo-EBLUP has slightly larger standard errors. FH-HT and FH-HA perform similarly, but FH-HA has less variation than FH-HT. FH-SRS performs poorly as expected under informative sampling.

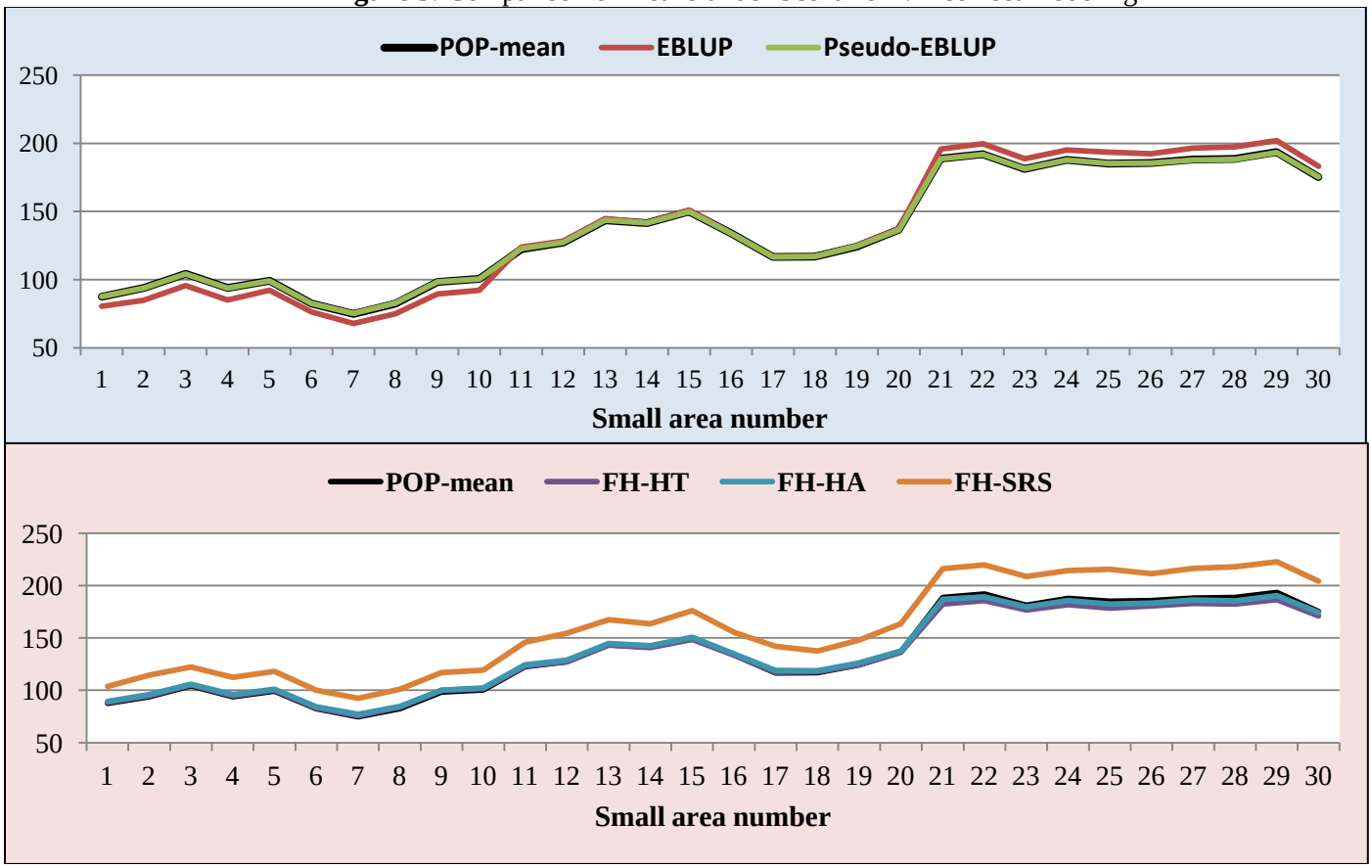
Figure 2: Comparison of standard errors of model-based estimators under Scenario I: correct modeling



We now compare the estimates under scenario II: incorrect modeling. Figure 3 compares the means. It is clear that pseudo-EBLUP is better than EBLUP under incorrect modeling. EBLUP leads to severe bias. Both the FH-HT and FH-

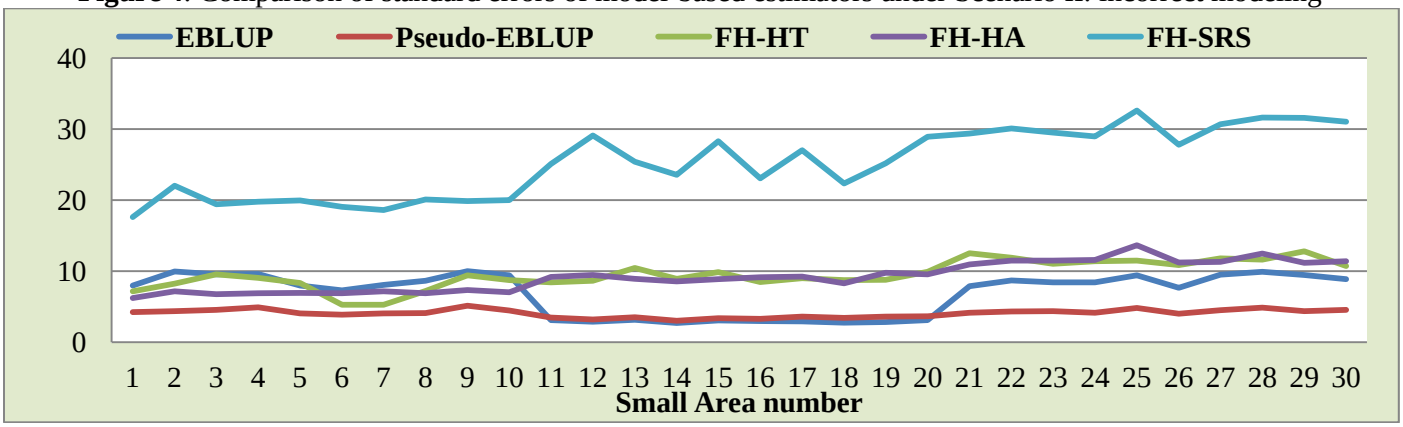
HA perform very well, and FH-SRS performs poorly with large bias. Figure 3 shows that using survey weights in the modeling is very important when the unit level model is incorrectly specified.

Figure 3: Comparison of means under Scenario II: incorrect modeling



For standard errors, from Figure 4 the pseudo-EBLUP performs the best under incorrect modeling. EBLUP has very large standard errors in model misspecification part. Again FH-HT and FH-HA perform similarly, and FH-SRS performs poorly. Thus pseudo-EBLUP performs the best in terms of standard errors under model misspecification.

Figure 4: Comparison of standard errors of model-based estimators under Scenario II: incorrect modeling



We now compare the confidence intervals. Figure 5 compares the confidence interval coverage rates under scenario I: correct modeling. The correlation between the selection probabilities and sampling units is presented as well to show the strength of informativeness of the PPS sampling. Figure 5 shows that, when the model is correct, the coverage rates for EBLUP, pseudo-EBLUP, FH-HT and FH-HA are quite stable under both informative and non-informative sampling, FH-HT has slightly lower coverage rate when the sampling is non-informative, whereas the coverage rate for FH-SRS is very poor (only about 25%) under informative sampling and increases to 95% under non-informative sampling.

Figure 5: Comparison of confidence intervals under Scenario I: correct modeling

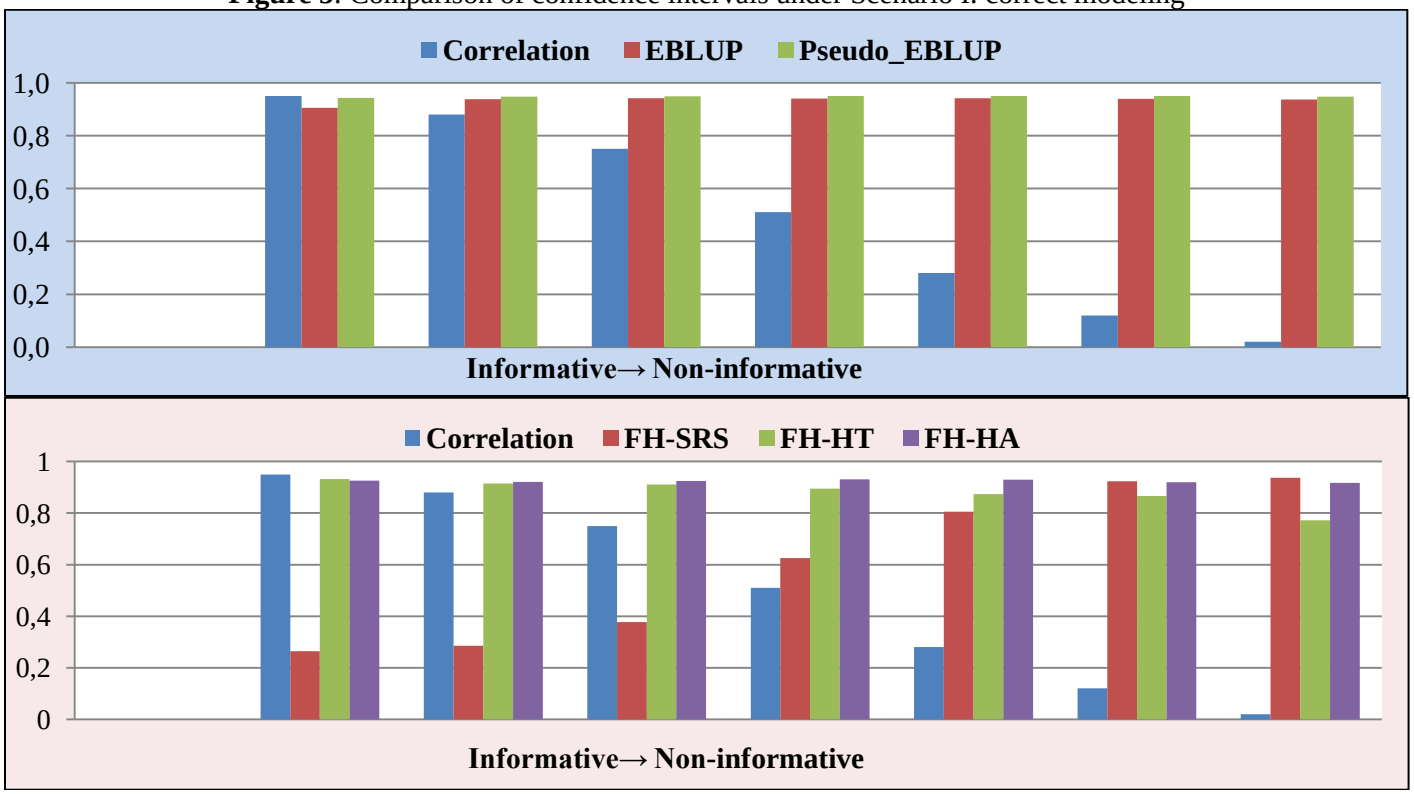
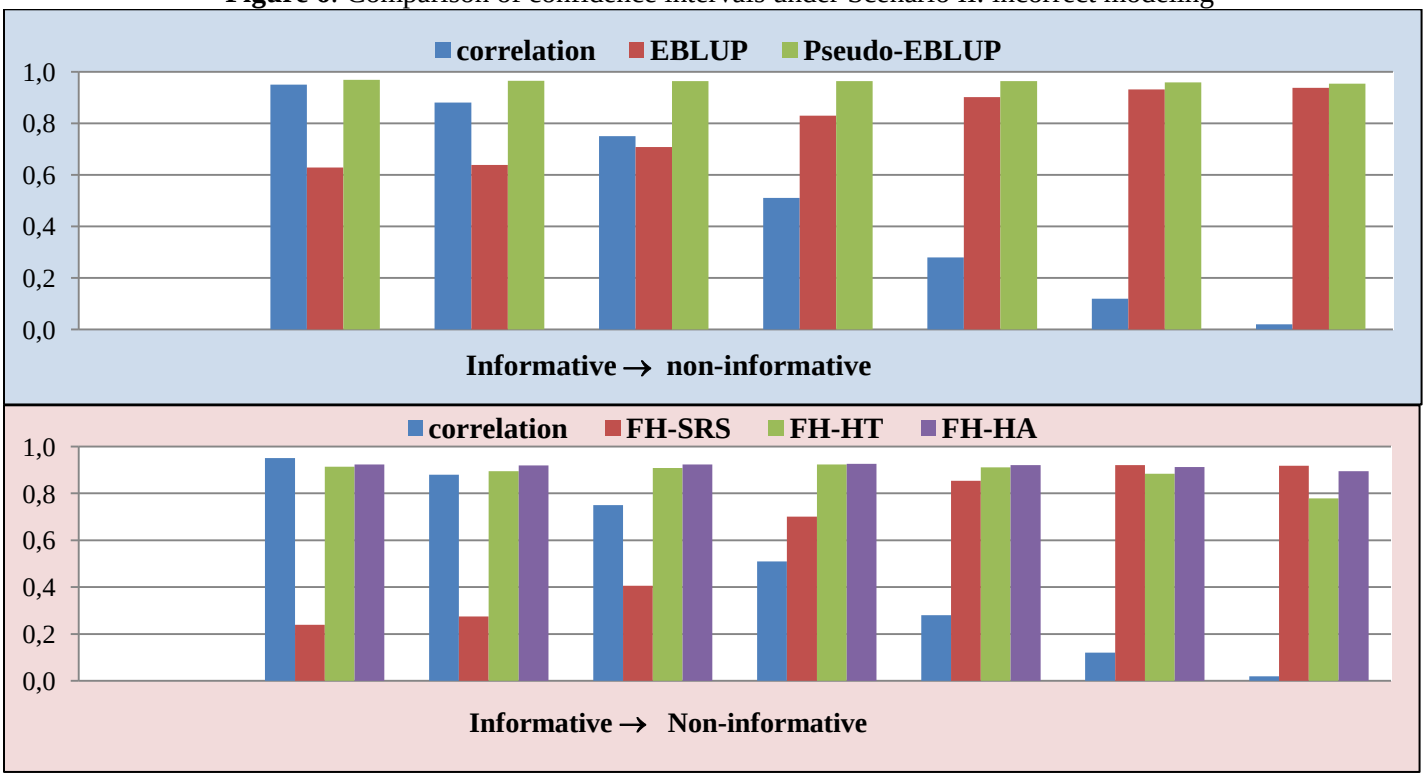


Figure 6 presents the coverage rates under scenario II: incorrect modeling. Figure 6 shows that the EBLUP has poor coverage rate (as low as 62%) under informative sampling, whereas the pseudo-EBLUP has very table and high coverage rates (all around 95%) under both the informative and non-informative sampling. For the FH estimators, FH-HT and FH-HA again perform well and have stable and high coverage rates, especially for the FH-HA. FH-HT has slightly lower coverage rate when the sampling is very non-informative. As expected, FH-SRS performs poorly when the sampling is informative

Figure 6: Comparison of confidence intervals under Scenario II: incorrect modeling



5. CONCLUSION

In this paper, we have compared the performance of the estimators based on the unit level nested error regression model and the area level Fay-Herriot model through a design-based simulation study. Overall the Pseudo-EBLUP estimator performs the best in terms of bias and coverage rate under both the informative and non-informative sampling. In practice, we suggest to construct the pseudo-EBLUP estimators using the survey weights and unit level observations. For area level models, FH-HA performs better than FH-HT. FH-SRS performs poorly. Thus we suggest to construct the weighted HA estimators and then apply the Fay-Herriot model to obtain the corresponding model-based estimators.

REFERENCES

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

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Datta, G. S. & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistics Sinica*, 10, 613–627.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 268-277.

Rivest, L.P. and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. Proceedings of the International Conference on Recent Advances in Survey Sampling, Ed. J.N.K. Rao.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

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

Torabi M, and Rao JNK (2010). The mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38, 598-608.

Wang, J. and Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. and Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.