Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comparison of unit level and area level small area estimators

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Michael A. Hidiroglou and Yong You

Release date: June 22, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Vol. 42, No. 1, pp. 41-61
Statistics Canada, Catalogue No. 12-001-X

Comparison of unit level and area level small area estimators

Michael A. Hidiroglou and Yong You[1]

Abstract

In this paper, we compare the EBLUP and pseudo-EBLUP estimators for small area estimation under the nested error regression model and three area level model-based estimators using the Fay-Herriot model. We conduct a design-based simulation study to compare the model-based estimators for unit level and area level models under informative and non-informative sampling. In particular, we are interested in the confidence interval coverage rate of the unit level and area level estimators. We also compare the estimators if the model has been misspecified. Our simulation results show that estimators based on the unit level model perform better than those based on the area level. The pseudo-EBLUP estimator is the best among unit level and area level estimators.

Key Words: Confidence interval; Design consistency; Fay-Herriot model; Informative sampling; Model misspecification; Nested error regression model; Relative root mean squared error (RRMSE); Survey weight.

1 Introduction

Model-based small area estimators have been widely used in practice to provide reliable indirect estimates for small areas in recent years. The model-based estimators are based on explicit models that provide a link to related small areas through supplementary data such as census and administrative records. Small area models can be classified into two broad types: (i) Unit level models that relate the unit values of the study variable to unit-specific auxiliary variables and (ii) Area level models that relate direct estimators of the study variable of the small area to the corresponding area-specific auxiliary variables. In general, area level models are used to improve the direct estimators if unit level data are not available. The sampling set-up is as in Rao (2003). That is, a universe U of size N is split into m non-overlapping small areas U_i of size N_i, where $i = 1, \ldots, m.$ Sampling is carried out in each small area using a probabilistic mechanism, resulting in samples s_i of size n_i. The selection probabilities associated with each element $j = 1, \ldots, n_i$ selected in sample s_i is denoted as p_{ij}. The resulting design weights are given by $w_{ij} = n_i^{-1} p_{ij}^{-1}$. In practice, these weights can be adjusted to account for non-response and/or auxiliary information. The resulting weights are known as the survey weights. In this paper, we assume full response to the survey, and no adjustment to the auxiliary data. Direct area level estimates are obtained for each area using the survey weights and unit observations from the area. The survey design can be incorporated into small area models in different ways. In the area level case, direct design-based estimators are modeled directly and the survey variance of the associated direct estimator is introduced into the model via the design-based errors. In the case of the unit level, the observations can be weighted using 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, that is, informativeness of the sampling process. In this paper, we compare, via a simulation study, the impact of model misspecification and the informativeness of the sampling design for two basic small area procedures based on unit and area levels in terms of bias, estimated mean squared error and confidence

[1] Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. E-mail: hidirog@yahoo.ca; Yong You, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. E-mail: yong.you@canada.ca.

interval coverage rates. A sampling design is informative if the selection probabilities p_{ij} are related to the variable of interest y_{ij} even after conditioning on the covariates $\mathbf{x}_{ij}$. In such cases, we have informative sampling in the sense that the population model no longer holds for the sample. Pfeffermann and Sverchkov (2007) accounted for this possibility by adjusting the small area procedures. Verret, Rao and Hidiroglou (2015) simplified the procedure. In this paper, we do not adjust the small area procedures for informativeness, but study their impact.

The paper is structured as follows. The point estimators and associated mean squared error estimators for the unit level and area models are described in Section 2 and in Section 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 (probability proportional to size with replacement) sampling scheme by varying the following two factors: (a) the assumed model is correct or incorrect; and (b) design informativeness varies from being non-significant to being very significant. In Section 5, we give an example using data from Battese, Harter and Fuller (1988) that compares the unit level and area level estimates. Finally, conclusions resulting from this work are presented in Section 6.

2 Unit level model

A basic unit level model for small area estimation is the nested error regression model (Battese et al. 1988) given by $y_{ij} = \mathbf{x}'_{ij}\beta + v_i + e_{ij}, \quad j = 1, \ldots, N_i, i = 1, \ldots, m,$ where y_{ij} is the variable of interest for the j^{th} population unit in the i^{th} small area, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is a $p \times 1$ vector of auxiliary variables, with $x_{ij1} = 1,$ $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ 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 independent and identically distributed $(i.i.d.)$ $N(0, \sigma_v^2)$ and independent of the unit errors e_{ij}, which are assumed to be $i.i.d.$ $N(0, \sigma_e^2)$. Assuming that N_i is large, the parameter of interest is the mean for the i^{th} area, $\bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, which may be approximated by

$$\theta_i = \bar{\mathbf{X}}'_i\boldsymbol{\beta} + v_i, \tag{2.1}$$

where $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij} \big/ N_i$ is the vector of known population means of the $\mathbf{x}_{ij}$ for the i^{th} area. We assume that samples are drawn independently within each small area according to a specified sampling design. Under non-informative sampling, the sample data $(y_{ij}, \mathbf{x}_{ij})$ are assumed to obey the population model, i.e.,

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{2.2}$$

where w_{ij} is the basic design weight associated with unit (i, j), 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, $\theta_i = \bar{\mathbf{X}}'_i\boldsymbol{\beta} + v_i$, based on the nested error regression model (2.2) is given by

$$\tilde{\theta}_i = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \tilde{\boldsymbol{\beta}}, \tag{2.3}$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \big/ n_i$, $r_i = \sigma_v^2 \big/ \left(\sigma_v^2 + \sigma_e^2 / n_i\right)$, and

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i{}' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \bar{\mathbf{x}}_i{}' \mathbf{V}_i^{-1} y_i\right) \equiv \tilde{\boldsymbol{\beta}}\left(\sigma_e^2, \sigma_v^2\right), \tag{2.4}$$

with $\mathbf{x}_i' = \left(x_{i1}, \ldots, x_{in_i}\right)$, $\mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}'$, $y_i = \left(y_{i1}, \ldots, y_{in_i}\right)'$, $i = 1, \ldots, m.$ Both $\tilde{\theta}_i$ and $\tilde{\boldsymbol{\beta}}$ depend on the unknown variance parameters σ_e^2 and σ_v^2. The method of fitting constant can be used to estimate σ_e^2 and σ_v^2, and the resulting estimators are $\hat{\sigma}_e^2 = (n - m - p + 1)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$, and $\hat{\sigma}_v^2 = \max\left(\tilde{\sigma}_v^2, 0\right)$, where $\tilde{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n - p)\hat{\sigma}_e^2\right]$, $n_* = n - \mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{m} n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'\right]$, $\mathbf{X}' = \left(x_1', \ldots, x_m'\right)$, $n = \sum_{i=1}^{m} n_i$.

The residuals $\left\{\hat{\varepsilon}_{ij}\right\}$ are obtained from the ordinary least squares (OLS) regression of $y_{ij} - \bar{y}_i$ on $\left\{\mathbf{x}_{ij1} - \bar{\mathbf{x}}_{i\cdot 1}, \ldots, \mathbf{x}_{ijp} - \bar{\mathbf{x}}_{i\cdot p}\right\}$ and $\left\{\hat{u}_{ij}\right\}$ are the residuals from the OLS regression of y_{ij} on $\left\{x_{ij1}, \ldots, x_{ijp}\right\}$. See Rao (2003), page 138 for more details.

Replacing σ_e^2 and σ_v^2 by estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ in equation (2.3), we obtain the EBLUP estimator of small area mean θ_i as

$$\hat{\theta}_i^{\text{EBLUP}} = r_i \bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{r}_i \bar{\mathbf{x}}_i\right)' \hat{\boldsymbol{\beta}}, \tag{2.5}$$

where $\hat{r}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i\right)$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$. The mean squared error (MSE) of the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ is given by

$$\mathrm{MSE}\left(\hat{\theta}_i^{\text{EBLUP}}\right) \approx g_{1i}\left(\sigma_e^2, \sigma_v^2\right) + g_{2i}\left(\sigma_e^2, \sigma_v^2\right) + g_{3i}\left(\sigma_e^2, \sigma_v^2\right),$$

see Prasad and Rao (1990). The $g-$terms are

$$\begin{aligned} g_{1i}\left(\sigma_e^2, \sigma_v^2\right) &= (1 - r_i)\sigma_v^2, \\ g_{2i}\left(\sigma_e^2, \sigma_v^2\right) &= \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right)' \left(\sum_{i=1}^{m} \mathbf{x}_i' \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1} \left(\bar{\mathbf{X}}_i - r_i \bar{\mathbf{x}}_i\right) \end{aligned}$$

and

$$g_{3i}\left(\sigma_e^2, \sigma_v^2\right) = n_i^{-2} \left(\sigma_v^2 + \sigma_e^2 n_i^{-1}\right)^{-3} h\left(\sigma_e^2, \sigma_v^2\right),$$

where $h\left(\sigma_e^2, \sigma_v^2\right) = \sigma_e^4 V\left(\tilde{\sigma}_v^2\right) - 2\sigma_e^2 \sigma_v^2 \,\mathrm{cov}\left(\hat{\sigma}_e^2, \tilde{\sigma}_v^2\right) + \sigma_v^4 V\left(\hat{\sigma}_e^2\right)$. The variances and covariance of $\hat{\sigma}_e^2$ and $\tilde{\sigma}_v^2$ are given by

$$\begin{aligned} V\left(\hat{\sigma}_e^2\right) &= 2(n - m - p + 1)^{-1} \sigma_e^4 \\ V\left(\tilde{\sigma}_v^2\right) &= 2n_*^{-2} \left[(n - m - p + 1)^{-1} (m - 1)(n - p)\sigma_e^4 + 2n_* \sigma_e^2 \sigma_v^2 + n_{**} \sigma_v^4\right], \end{aligned}$$

and

$$\mathrm{cov}\left(\hat{\sigma}_e^2, \tilde{\sigma}_v^2\right) = -(m - 1) n_*^{-1} V\left(\hat{\sigma}_e^2\right),$$

where $n_{**} = \operatorname{tr}(\mathbf{Z}'\mathbf{M}\mathbf{Z})^2$, $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{Z} = \operatorname{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$.

A second-order unbiased estimator of the MSE (Prasad and Rao 1990) is given by

$$\operatorname{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right) = g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.6}$$

Note that the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ given by (2.5) depends on the unit level model (2.2). It is model-unbiased, but it is not design consistent unless the sample design is simple random sampling. If model (2.2) does not hold for the sampled data, then the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ may be biased, that is, additional bias will be present in the EBLUP estimator due to model misspecification.

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 (2.2) to achieve design consistency. Let w_{ij} be the weights associated with each unit (i, j). A direct design-based estimator of the small area mean is given by

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}} = \sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}, \tag{2.7}$$

where $\tilde{w}_{ij} = w_{ij} \big/ \sum_{j=1}^{n_i} w_{ij} = w_{ij} / w_{i.}$ and $\sum_{j=1}^{n_i} \tilde{w}_{ij} = 1$. The weighted estimator $\bar{y}_{iw}$ is also known as the weighted Hájek estimator. By combining the direct estimator (2.7) and the unit level model (2.2), we can obtain the following aggregated (survey-weighted) area level model

$$\bar{y}_{iw} = \bar{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \bar{e}_{iw}, \quad i = 1, \ldots, m, \tag{2.8}$$

where $\bar{e}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} e_{ij}$ with $E(\bar{e}_{iw}) = 0$, $V(\bar{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} \tilde{w}_{ij}^2 \equiv \delta_i^2$, and $\bar{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ij} \mathbf{x}_{ij}$. Note that the regression parameter $\boldsymbol{\beta}$ and the variance components σ_e^2 and σ_v^2 are unknown in model (2.8). Based on model (2.8), assuming that the parameters $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 are known, the BLUP estimator of θ_i is

$$\tilde{\theta}_{iw} = r_{iw}\bar{y}_{iw} + \left(\bar{\mathbf{X}}_i - r_{iw}\bar{\mathbf{x}}_{iw}\right)'\boldsymbol{\beta} = \tilde{\theta}_{iw}\left(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2\right), \tag{2.9}$$

where $r_{iw} = \sigma_v^2 \big/ \left(\sigma_v^2 + \delta_i^2\right)$. 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 of $\boldsymbol{\beta}$ as follows:

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

$\tilde{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\sigma_e^2, \sigma_v^2\right)$ depends on σ_e^2 and σ_v^2. Replacing σ_e^2 and σ_v^2 in $\tilde{\boldsymbol{\beta}}_w$ by the fitting of constant estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$ is obtained; See Rao (2003, page 149). Replacing $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 in (2.9) by $\hat{\boldsymbol{\beta}}_w$, $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, the pseudo-EBLUP estimator for the small area mean θ_i is given by

$$\hat{\theta}_i^{P-\text{EBLUP}} \triangleq \hat{\theta}_{iw} = \hat{r}_{iw}\bar{y}_{iw} + \left(\bar{\mathbf{X}}_i - \hat{r}_{iw}\bar{\mathbf{x}}_{iw}\right)' \hat{\boldsymbol{\beta}}_w. \tag{2.10}$$

As the sample size n_i becomes large, estimator $\hat{\theta}_i^{P-\text{EBLUP}}$ becomes design-consistent. It also has a self-benchmarking property when the weights w_{ij} are calibrated to agree with the known population total. That is, if $\sum_{j=1}^{n_i} w_{ij} = N_i$, $\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\text{EBLUP}}$ is equal to the direct regression estimator of the overall total,

$$\sum_{i=1}^{m} N_i \hat{\theta}_i^{P-\text{EBLUP}} = \hat{Y}_w + \left(\mathbf{X} - \hat{\mathbf{X}}_w\right)' \hat{\boldsymbol{\beta}}_w,$$

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

The MSE of $\hat{\theta}_i^{P-\text{EBLUP}}$ is given by

$$\text{MSE}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) \approx g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) + g_{3iw}\left(\sigma_e^2, \sigma_v^2\right),$$

where $g_{1iw}\left(\sigma_e^2, \sigma_v^2\right) = (1 - r_{iw})\sigma_v^2$, $g_{2iw}\left(\sigma_e^2, \sigma_v^2\right) = \left(\bar{X}_i - r_{iw}\bar{x}_{iw}\right)'\Phi_w\left(\bar{X}_i - r_{iw}\bar{x}_{iw}\right)$. The term Φ_w is

$$\begin{aligned}\Phi_w &\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}z'_{ij}\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{z}_{ij}\mathbf{z}'_{ij}\right)\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}'_{ij}\right)^{-1}\right]'\sigma_e^2 \\ &+ \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}'_{ij}\right)^{-1}\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)'\right]\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}x_{ij}\mathbf{z}_{ij}\right)^{-1}\right]'\sigma_v^2,\end{aligned}$$

where $\mathbf{z}_{ij} = w_{ij}\left(\mathbf{x}_{ij} - r_{iw}\bar{\mathbf{x}}_{iw}\right)$, $g_{3iw}\left(\sigma_e^2, \sigma_v^2\right) = r_{iw}(1 - r_{iw})^2\sigma_e^{-4}\sigma_v^{-2}h\left(\sigma_e^2, \sigma_v^2\right)$. $h\left(\sigma_e^2, \sigma_v^2\right)$ is the same function as in the MSE for the EBLUP estimator given in Section 2.1. A nearly second-order unbiased estimator of the MSE can be obtained as

$$\text{mse}\left(\hat{\theta}_i^{P-\text{EBLUP}}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right). \tag{2.11}$$

(See Rao 2003, page 150 and You and Rao 2002, page 435). Note that the MSE estimator (2.11) ignores the cross-product terms. Torabi and Rao (2010) obtained the second-order correct MSE estimator including the cross-product terms using linearization and bootstrap methods. There are two cross-product terms. The first one is simple and has a closed form. Although the linearization method performs well, the explicit form for the second cross-product term is very lengthy: furthermore, the formulas based on the linearization procedure are not provided in Torabi and Rao (2010). The bootstrap method always underestimates the true MSE. A double bootstrap method needs to be applied to get an unbiased estimator of the MSE and is computationally extensive. The MSE estimator (2.11) behaves like the linearization estimator of Torabi and Rao (2010) when the variation of the survey weights is small. In the case of self-weighting within areas, one of the cross-product term is zero and the other term is of order $o\left(m^{-1}\right)$. Hence, the MSE estimator (2.11) is nearly unbiased; more discussion is provided in Torabi and Rao (2010). It is for these reasons that these cross-product terms were not included in the MSE estimator given by (2.11) in our study.

Note that under model (2.2) the pseudo-EBLUP estimator $\hat{\theta}_i^{P-\text{EBLUP}}$ is slightly less efficient than the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$. However, the pseudo-EBLUP estimator is design consistent and is therefore more robust to 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 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 $\hat{\theta}_i^{\text{DIR}}$. The direct survey estimator is design unbiased for the small area parameter θ_i. The sampling model is given by

$$\hat{\theta}_i^{\text{DIR}} = \theta_i + e_i,\ i = 1, \ldots, m, \tag{3.1}$$

where the e_i is the sampling error associated with the direct estimator $\hat{\theta}_i^{\text{DIR}}$ 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 $E(e_i) = 0$ and sampling variance $\text{var}(e_i) = \sigma_i^2$. The linking model is obtained by assuming that the small area parameter of interest θ_i is related to area level auxiliary variables $\mathbf{z}_i = (\text{z}_{i1}, \ldots, \text{z}_{ip})'$ through the following linear regression model

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i,\ i = 1, \ldots, m, \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, and the v_i's are area-specific random effects assumed to be $i.i.d.$ with $E(v_i) = 0$ and $\text{var}(v_i) = \sigma_v^2$. The assumption of normality is generally also made, even though it is more difficult to justify the assumption. This assumption is needed to obtain the MSE estimation. The model variance σ_v^2 is unknown and needs to be estimated from the data. The area level random effect v_i capture the unstructured heterogeneity among areas that is not explained by the sampling variances. Combining models (3.1) and (3.2) leads to a linear mixed area level model given by

$$\hat{\theta}_i^{\text{DIR}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i. \tag{3.3}$$

Model (3.3) 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 (3.3). This is a very strong assumption. Generally smoothed estimators of the sampling variances are used in the Fay-Herriot model and then σ_i^2's are treated as known. However, if direct estimators of sampling variances are used in the Fay-Herriot model, an extra term needs to be added to the MSE estimator to account for the extra variation (Wang and Fuller 2003).

Assuming that 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

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i^{\text{DIR}} + (1 - \gamma_i)\, \mathbf{z}_i'\tilde{\boldsymbol{\beta}}_{\text{WLS}}, \tag{3.4}$$

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

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

There are several methods available to estimate the unknown model variance σ_v^2; You (2010) provides a review of these methods. We chose the restricted maximum likelihood (REML) obtained by Cressie (1992) to estimate the model variance under the Fay-Herriot model. Using the scoring algorithm, the REML estimator $\hat{\sigma}_v^2$ is obtained as

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

where $I_R\left(\sigma_v^2\right) = 1/2\,\text{tr}[\mathbf{PP}]$, and $S_R\left(\sigma_v^2\right) = 1/2\,\mathbf{y}'\mathbf{PPy} - 1/2\,\text{tr}[\mathbf{P}]$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{V}^{-1}$. Using a guessing value for $\sigma_v^{2(1)}$ as the starting value, the algorithm converges very fast.

Replacing σ_v^2 in equation (3.4) by the REML estimator $\hat{\sigma}_v^2$, we obtain the EBLUP of the small area parameter θ_i based on the Fay-Herriot model as

$$\hat{\theta}_i^{\text{FH}} = \hat{\gamma}_i\hat{\theta}_i^{\text{DIR}} + (1 - \hat{\gamma}_i)\,\mathbf{z}_i'\hat{\boldsymbol{\beta}}_{\text{WLS}}, \tag{3.5}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \sigma_i^2\right)$. The MSE estimator of $\hat{\theta}_i^{\text{FH}}$ is given by (see Rao 2003)

$$\text{mse}\left(\hat{\theta}_i^{\text{FH}}\right) = g_{1i} + g_{2i} + 2g_{3i}, \tag{3.6}$$

where g_{1i} is the leading term, g_{2i} accounts for the variability due to estimation of the regression parameter β, and g_{3i} is due to the estimation of the model variance. These $g-$terms are defined as follow:

$$g_{1i} = \hat{\gamma}_i\sigma_i^2, g_{2i} = (1 - \hat{\gamma}_i)^2\,\mathbf{z}_i'\text{var}\left(\hat{\boldsymbol{\beta}}_{\text{WLS}}\right)\mathbf{z}_i = \hat{\sigma}_v^2\,(1 - \hat{\gamma}_i)^2\,\mathbf{z}_i'\left(\sum_{i=1}^{m}\hat{\gamma}_i\mathbf{z}_i\mathbf{z}_i'\right)^{-1}\mathbf{z}_i$$

and $g_{3i} = \left(\sigma_i^2\right)^2\left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-3}\text{var}\left(\hat{\sigma}_v^2\right)$.

The estimated variance of $\hat{\sigma}_v^2$ is given by $\text{var}\left(\hat{\sigma}_v^2\right) = 2\left(\sum_{i=1}^{m}\left(\hat{\sigma}_v^2 + \sigma_i^2\right)^{-2}\right)^{-1}$; see Datta and Lahiri (2000).

Up to now we have assumed that the sampling variance σ_i^2 is assumed known in the Fay-Herriot model (3.3). This is a very strong assumption. Usually a direct survey estimator, say s_i^2, of the sampling variance σ_i^2 is available. As these estimated variances can be quite variable, they are smoothed using external models and generalized variance functions: these smoothed variances are denoted as $\tilde{s}_i^2$. The smoothed sampling variance estimates $\tilde{s}_i^2$ are used in the Fay-Herriot model and treated as known. The associated $\text{mse}\left(\hat{\theta}_i^{\text{FH}}\right)$ is obtained by replacing σ_i^2 by $\tilde{s}_i^2$ in equation (3.6). 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 (3.6), and this term, denoted as g_{4i}, is given by

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

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

To apply the Fay-Herriot model, we need to obtain area level direct estimates and the corresponding sampling variance estimates as input values for the Fay-Herriot model. We consider three area level direct estimators; namely, the direct sample mean estimator assuming simple random sampling (SRS), the Horvitz-Thompson estimator (HT), and the weighted Hájek estimator (HA). The weighted Hájek estimator is also used in the pseudo-EBLUP estimator for the unit level model denoted as $\bar{y}_{iw}$ in equation (2.7). Table 3.1 presents these three area level direct estimators and the corresponding sampling variance estimators.

Table 3.1
Area level direct estimators and sampling variances

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

These area level estimators are used as input values into the Fay-Herriot model. Correspondingly, the three area level model-based estimators are denoted as: FH-SRS, FH-HT, and FH-HA. That is, we replace $\hat{\theta}_i^{\text{DIR}}$ by $\hat{\theta}_i^{\text{SRS}}$, $\hat{\theta}_i^{\text{HT}}$ or $\hat{\theta}_i^{\text{HA}}$ in (3.5) and obtain the corresponding model-based estimator $\hat{\theta}_i^{\text{FH-SRS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$. The SRS direct estimator $\hat{\theta}_i^{\text{SRS}}$ ignores the sample design and is not design consistent, unless the sample design is based on simple random sampling. Note that $\hat{\theta}_i^{\text{HT}}$ and $\hat{\theta}_i^{\text{HA}}$ are design consistent estimators. It follows that the corresponding model-based estimators $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$ are design consistent as the sample size increases. Furthermore, this means that these estimators are robust to model misspecification.

In the next section, we compare the unit level model with the Fay-Herriot model through a simulation study. The statistics used for these comparisons are bias, relative root MSE and confidence intervals of the model-based estimators.

4 Simulation study

4.1 Data generation

To compare the unit level and area level small area estimators, we conducted a design-based simulation study. Following the simulation setup of You, Rao and Kovacevic (2003), we created two finite populations. Each finite population had $m = 30$ areas, and each area consisted of $N_i = 200$ population units. Each finite population was generated using the unit level model $y_{ij} = \beta_0 + x_{1ij}\beta_1 + v_i + e_{ij}$. The auxiliary variable x_{1ij} was generated from an exponential distribution with mean 4 and variance 8, and the random components were generated from the normal distribution with $v_i \sim N\left(0, \sigma_v^2\right)$, $e_{ij} \sim N\left(0, \sigma_e^2\right)$, where $\sigma_v^2 = 100$ and $\sigma_e^2 = 225$. For the first population, the regression fixed effects were set as $\beta_0 = 50$, $\beta_1 = 10$ for all 30 areas. For the second population, different fixed effects values were used: $\beta_0 = 50$, $\beta_1 = 10$ for areas $m = 1, \ldots, 10$; $\beta_0 = 75$, $\beta_1 = 15$ for areas $m = 11, \ldots, 20$; $\beta_0 = 100$, $\beta_1 = 20$ for areas $m = 21, \ldots, 30$. We had three different means for the fixed effects $\beta_0 + x_{1ij}\beta_1$ in the second population, whereas we only had one in the first population. PPSWR samples within each area were drawn independently from each constructed population. PPSWR sampling was implemented as follows: We first defined a size measure z_{ij} for a given unit (i, j). Using these z_{ij} values, we computed selection probabilities $p_{ij} = z_{ij} \big/ \sum_j z_{ij}$ for each unit (i, j) and used them to select PPSWR samples of equal size $n_i = n$. Within each generated population, we selected samples of size $n = 10$ and 30. The basic design weight is given by $w_{ij} = n_i^{-1} p_{ij}^{-1}$, so that the standardized weight is $\tilde{w}_{ij} = p_{ij}^{-1} \big/ \sum_j p_{ij}^{-1}$. We chose the size measure z_{ij} as a linear combination of the auxiliary variable x_{1ij} and data generated from an exponential distribution with mean 4 and variance 16. The correlation coefficient ρ between y_{ij} and the selection probability p_{ij} within each area varied between 0.02 and 0.95. The range of the p_{ij}'s corresponds to non-informative selection $(\rho = 0.02)$ to strongly informative selection $(\rho = 0.95)$ of the PPSWR samples. The sampling is non-informative when the correlation coefficient between y_{ij} and the selection probability p_{ij} is very weak, implying that the sample and the population model coincide. If the selection probability p_{ij} is strongly correlated with the observation y_{ij}, we have informative sampling, and the population model may no longer holds for the sample. For each population, the PPSWR sampling process was repeated $R = 3{,}000$ times. As in Prasad and Rao (1990), the simulation study is design-based as both the populations were generated only once, and repeated samples were generated from the same population.

For unit level modeling, we fitted the nested error regression model to the PPSWR sampling data generated from each population. We obtained the corresponding EBLUP and pseudo-EBLUP estimates and related MSE estimates using the formulas given in Section 2. We then constructed the confidence interval estimates using the squared root of the MSE estimates; details are given in Section 4.2.3. For area level modeling, we first obtained the direct area level estimates $\hat{\theta}_i^{\text{SRS}}$, $\hat{\theta}_i^{\text{HT}}$ and $\hat{\theta}_i^{\text{HA}}$ as well as the corresponding sampling variances. We applied the Fay-Herriot model and obtained the model-based estimators $\hat{\theta}_i^{\text{FH-SRS}}$, $\hat{\theta}_i^{\text{FH-HT}}$ and $\hat{\theta}_i^{\text{FH-HA}}$. The population mean of the auxiliary variable x_{1ij} within each area was used in the Fay-Herriot model as the auxiliary variable. The g_{4i} was added to the MSE estimator to account for the use of unsmoothed sampling variances in the Fay-Herriot model. The corresponding confidence intervals were obtained similarly for the unit level EBLUP and pseudo-EBLUP estimators.

For both unit level and area level model fitting, we used the following two scenarios: Scenario I: correct modeling, where the data were generated from the first population and the fitting models were unit level model (2.2) and area level model (3.3) with common $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Scenario II: incorrect modeling, where the data were 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 $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Note that under scenario I the sampling is noninformative when the correct unit level (2.2) is fitted to the sample data to obtain the EBLUP estimator: this is true for any correlation coefficient ρ between y_{ij} and p_{ij}.

4.2 Results

In this section, we compare a number of statistics for the unit level and area level estimates under both scenario I (correct modeling) and scenario II (incorrect modeling).

4.2.1 Comparison within each small area

Figure 4.1 compares the population means with the unit level and area level estimates when $n = 10$ for scenario I. The results are based on a strongly informative sampling design where the correlation coefficient between y_{ij} and the selection probability p_{ij} is $\rho = 0.88$. The model-based estimates are based on the average of $R = 3{,}000$ simulation runs. It is clear from Figure 4.1 that the unit level estimators EBLUP (equation 2.5) and pseudo-EBLUP (equation 2.10) are almost unbiased. The results show that under correct modeling, the sampling is noninformative with respect to unit level model (2.2), and the EBLUP is unbiased. The area level estimator FH-SRS consistently overestimates the population mean, leading to a large bias. The area level estimator FH-HT generally underestimates the population mean and has slightly larger bias than the FH-HA estimator. For $n = 30$, we obtained similar results.

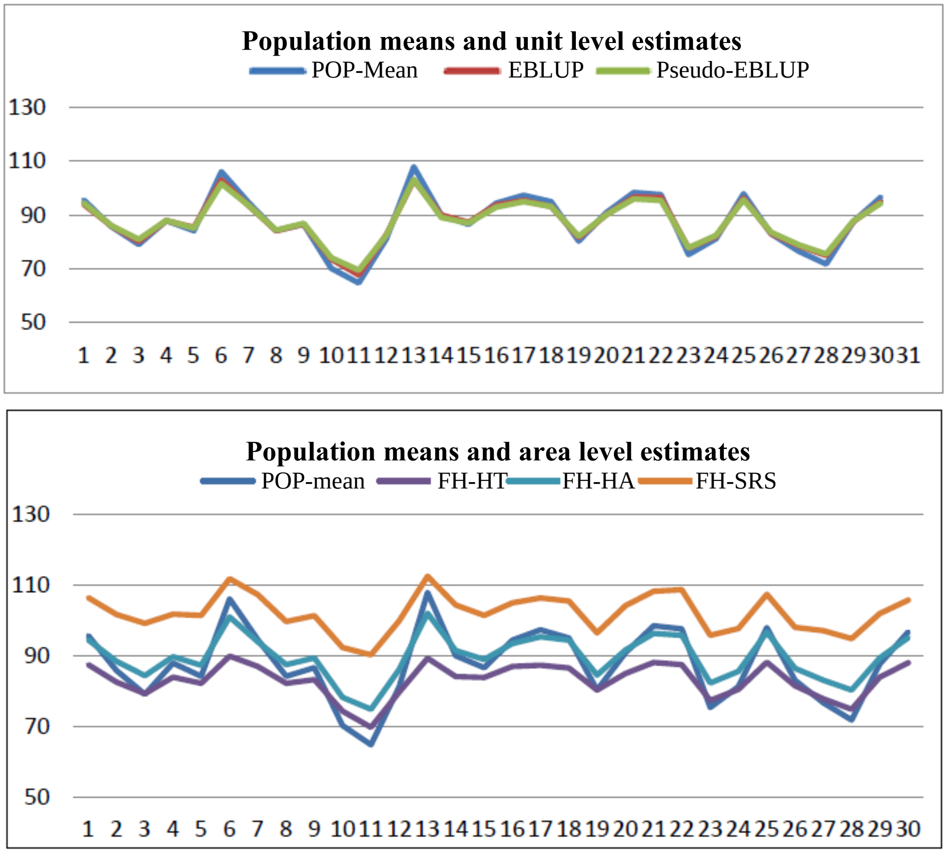

Figure 4.1 Comparison of means under scenario I: $n = 10$.

Figure 4.2 compares the average root mse for both unit level and area level estimators for scenario I when $n = 10$ and $n = 30$. The root mse's are the squared root of the estimated MSE's given in Sections 2 and 3 for the unit level and area level estimators. It is clear that EBLUP and pseudo-EBLUP have much smaller root mse's than the FH area level estimators for both $n = 10$ and $n = 30$. As expected (You and Rao 2002), EBLUP has the smallest root mse and pseudo-EBLUP has slightly larger root mse. For area level estimators, FH-SRS has large root mse and large variations. FH-HT and FH-HA have on average about the same root mse, but FH-HT is more variable than FH-HA as shown in both figures, particularly when sample size $n = 10$. When the sample size $n = 30$, the variability of the root mse's for FH-HT and FH-HA are substantially reduced, but it is clear that FH-HA is more stable than FH-HT.

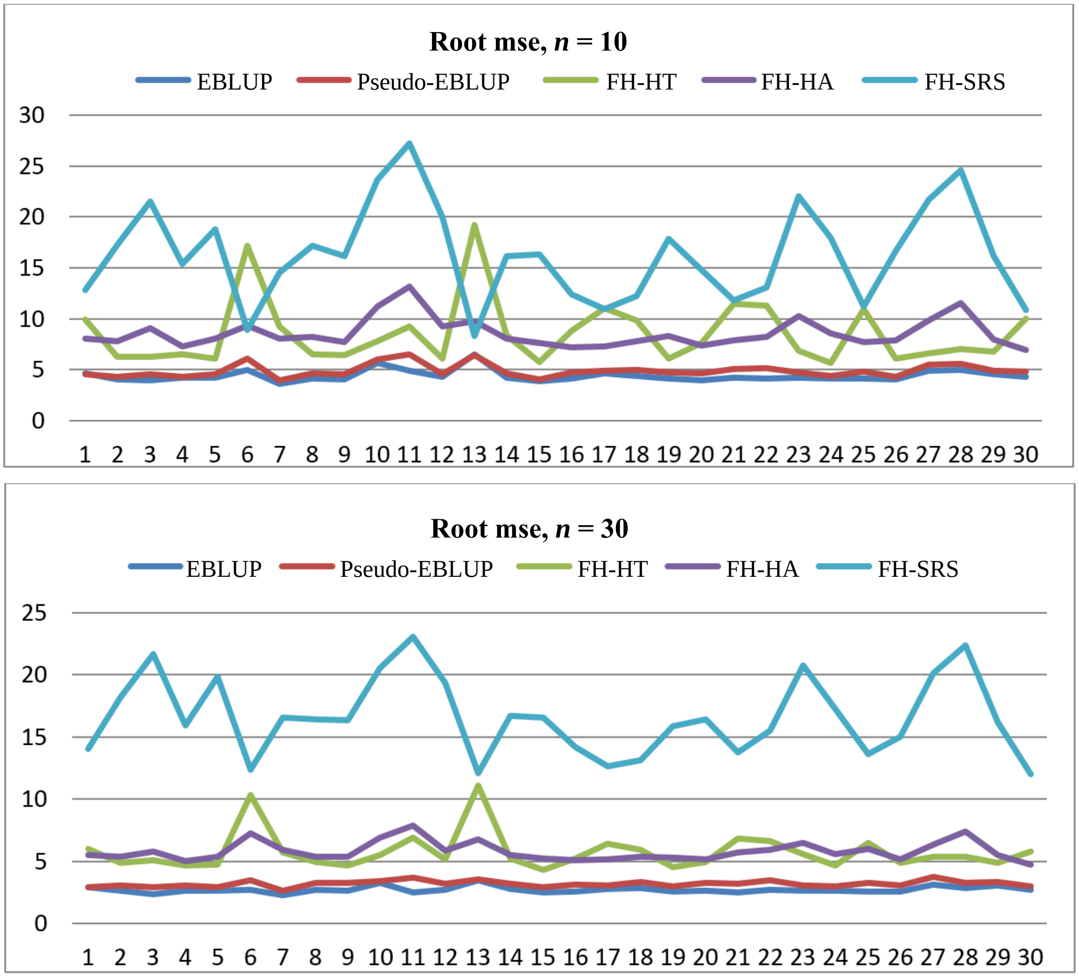

Figure 4.2 Comparison of root mse under scenario I: $n = 10$ and $n = 30$.

Figure 4.3 compares the unit level and area level estimates with the population means when $n = 10$ under scenario II. For unit level models, it is clear that EBLUP both underestimates and overestimates the population mean when the model is misspecified, whereas pseudo-EBLUP is unbiased (the pseudo-EBLUP estimates and population means overlap in Figure 4.3). For area level estimators, FH-SRS consistently overestimates the true means, while FH-HT has more underestimation than FH-HA as shown when the model is misspecified.

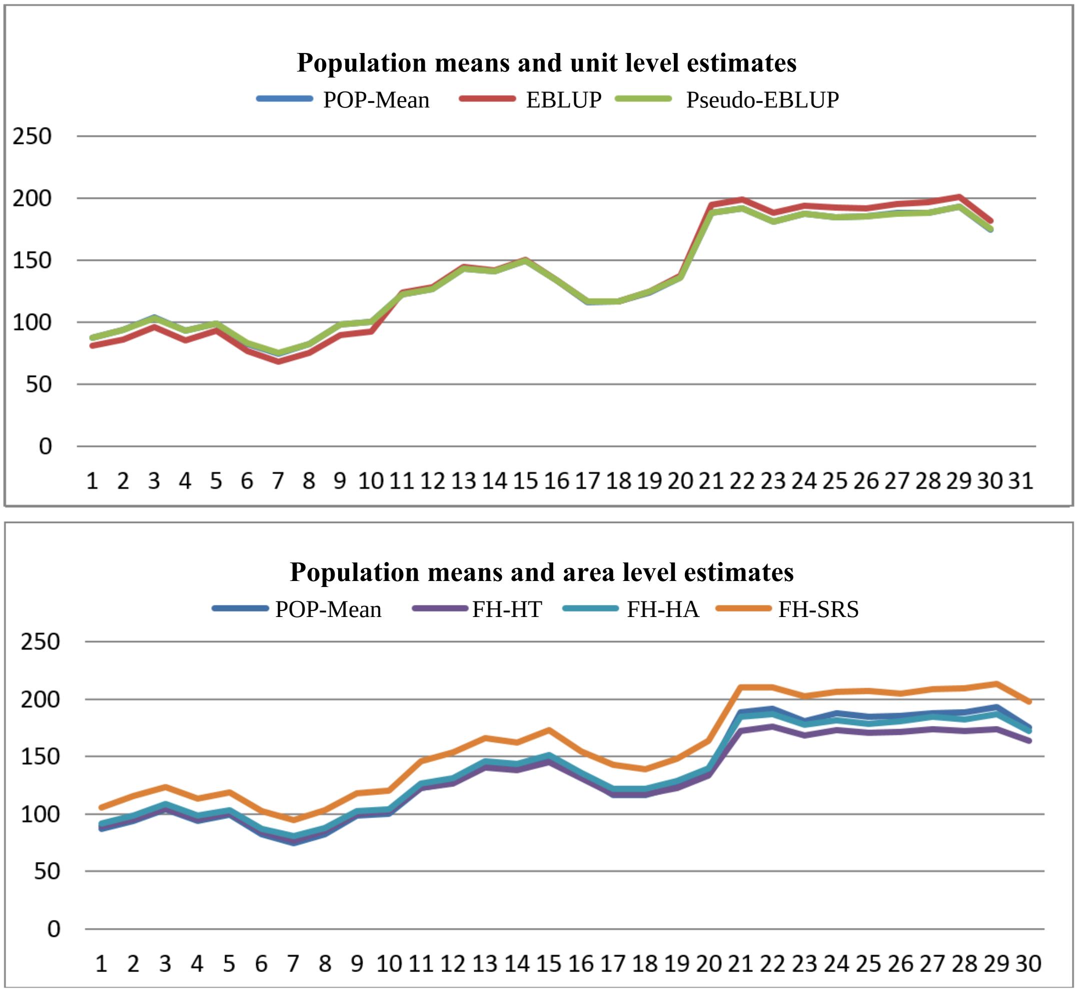

Figure 4.3 Comparison of means under scenario II: incorrect modeling, $n = 10$.

Figure 4.4 compares the root mse's of the unit level and area level estimators for both sample size $n = 10$ and $n = 30$ under incorrect modeling. From Figure 4.4, it can be seen that the pseudo-EBLUP estimator has the smallest root mse under incorrect modeling. EBLUP has very large root mse when the model is misspecified: that is, for areas 1 to 10 and areas 21 to 30, the average root mse is 10.01, whereas for pseudo-EBLUP, the corresponding root mse is 7.38 when the sample size $n = 10$. When the sample size $n = 30$, the average root mse is 8.85 for EBLUP, and only 4.38 for pseudo-EBLUP when the model is misspecified. In summary, the results show that the EBLUP estimator leads to biased estimates with large root mse under incorrect modeling.

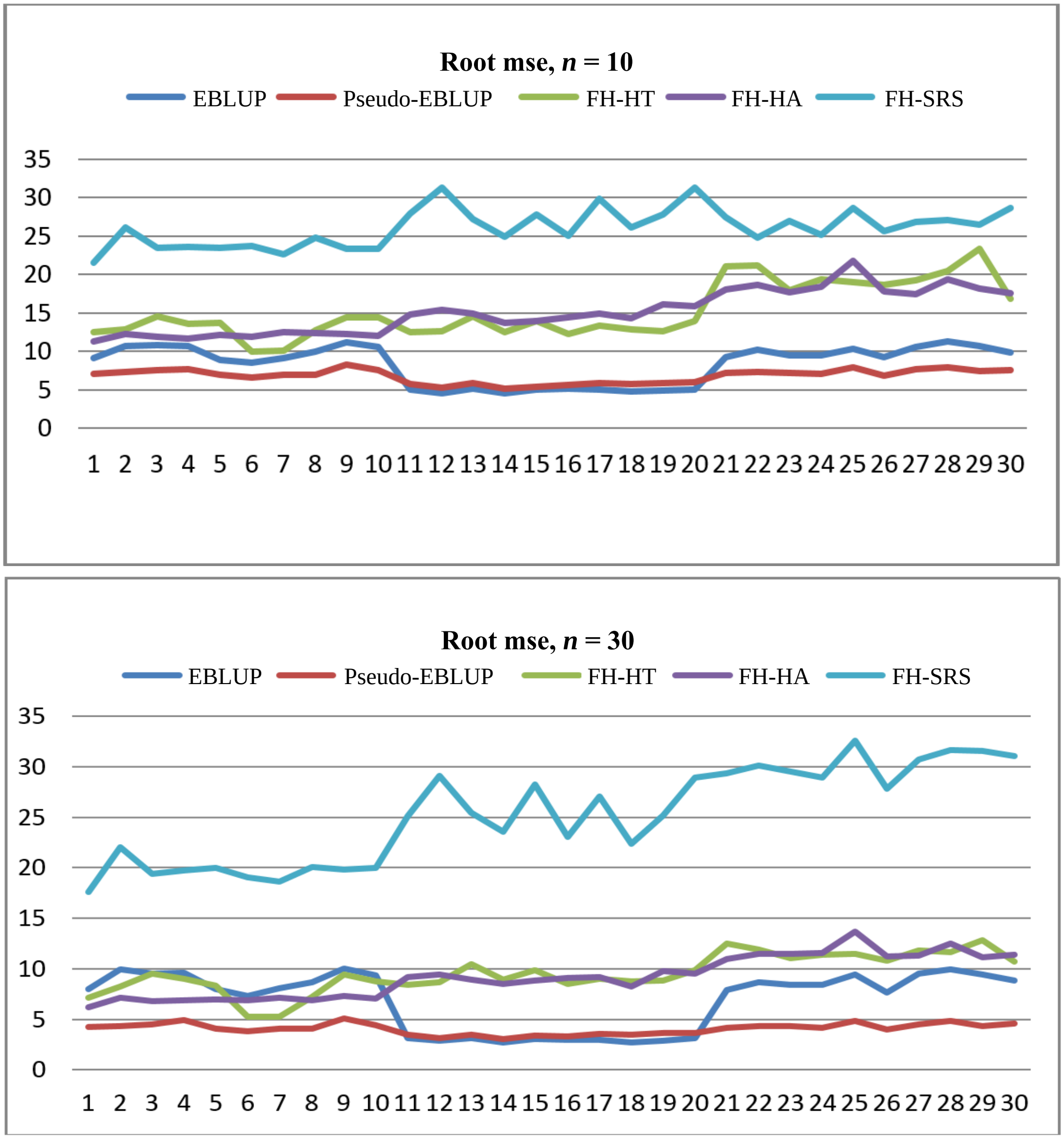

Figure 4.4 Comparison of root mse under scenario II: $n = 10$ and $n = 30$.

4.2.2 Comparison across small areas

To compare the estimators across areas, we considered the average absolute relative bias (ARB) for a specified estimator $\hat{\theta}_i$ of the simulated population mean $\bar{Y}_i$ as $\overline{\text{ARB}} = \left(\sum_{i=1}^{m} \text{ARB}_i\right) \Big/ m,$ where

$$\text{ARB}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)}{\bar{Y}_i} \right|,$$

and $\hat{\theta}_i^{(r)}$ is the estimate based on the r^{th} simulated sample, $R = 3{,}000, m = 30.$ Table 4.1 displays the percentage of the average absolute relative bias $\overline{\text{ARB}}$ of unit level and area level estimators over the 30 area for scenario I. The results are based on samples selected with sample sizes equal to 10 and 30 respectively within each area.

Table 4.1
Average absolute relative bias $\overline{\text{ARB}}\%$ for scenario I

Type	Estimator	*n* = 10	*n* = 30
Unit level	EBLUP	1.71	0.75
	Pseudo-EBLUP	2.14	0.86
Area level	FH-SRS	17.51	18.64
	FH-HT	6.02	3.12
	FH-HA	4.33	2.59

For unit level models, it is clear that if we use the correct model, the sample becomes noninformative with respect to unit level model (2.2), and both EBLUP and pseudo-EBLUP estimators are unbiased. The average absolute relative bias $\overline{\text{ARB}}$ for EBLUP is 1.71% when the sample size $n = 10$ and 0.75% when the sample size $n = 30.$ For pseudo-EBLUP, the $\overline{\text{ARB}}$ is 2.14% when $n = 10$ and 0.86% when $n = 30,$ respectively. Pseudo-EBLUP has slightly larger bias than EBLUP. For area level models, FH-SRS severely overestimates the means with the average $\overline{\text{ARB}}$ as large as 17.51% when $n = 10$ and 18.6% when $n = 30.$ Both area level estimators FH-HT and FH-HA lead to reasonable estimates: (i) The $\overline{\text{ARB}}$ for FH-HT is 6.02% when $n = 10$ and 3.12% when $n = 30;$ (ii) The $\overline{\text{ARB}}$ for FH-HA is 4.33% when $n = 10$ and 2.59% when $n = 30.$ The FH-HA estimator performs better than the FH-HT estimator. The absolute relative bias for the area level estimators is larger than the one associated with the unit level estimators.

Table 4.2 displays the $\overline{\text{ARB}}$ of the various estimators under scenario II. It is clear that pseudo-EBLUP has a much smaller $\overline{\text{ARB}}$ than EBLUP under incorrect modeling. The $\overline{\text{ARB}}$'s for EBLUP under incorrect modeling are 4.31% $(n = 10)$ and 4.52% $(n = 30)$ respectively. For pseudo-EBLUP, the average $\overline{\text{ARB}}$ is only 0.25% $(n = 10)$ and 0.12% $(n = 30).$ Both FH-HT and FH-HA perform very well. Their average $\overline{\text{ARB}}$'s are 3.91% and 3.48% respectively when $n = 10.$ These $\overline{\text{ARB}}$'s decrease to 1.51% and 1.47% when $n = 30.$ FH-SRS performs poorly. Both area level estimators FH-HT and FH-HA perform well and these estimators are also design consistent. Again, FH-HA is slightly better than FH-HT in terms of $\overline{\text{ARB}}.$ The results show that the use of survey weights in the unit level modeling is very important when the unit level model is incorrectly specified. The pseudo-EBLUP estimator leads to unbiased estimator even when the model is incorrectly specified. It is the best estimator when the model is incorrect.

Table 4.2
Average absolute relative bias $\overline{\text{ARB}}\%$ for scenario II

Type	Estimator	*n* = 10	*n* = 30
Unit level	EBLUP	4.31	4.52
	Pseudo-EBLUP	0.25	0.12
Area level	FH-SRS	17.11	17.87
	FH-HT	3.91	1.51
	FH-HA	3.48	1.47

We now compare the relative root MSE for all the estimators. In particular, we computed both the true simulation relative root MSE (RRMSE) and the estimated relative root MSE based on the MSE estimators. The average true simulation relative root MSE is computed as $\overline{\text{RRMSE}} = \left(\sum_{i=1}^{m} \text{RRMSE}_i\right) \Big/ m,$ where

$$\text{RRMSE}_i = \frac{\sqrt{\text{MSE}_i}}{\bar{Y}_i}, \text{ and } \text{MSE}_i = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \bar{Y}_i\right)^2.$$

The average estimated relative root MSE is computed as $\overline{\text{RRmse}} = \left(\sum_{i=1}^{m} \text{RRmse}_i\right) \big/ m,$ where

$$\text{RRmse}_i = \frac{\sqrt{\text{mse}_i}}{\hat{\theta}_i}, \text{ and } \text{mse}_i = \frac{1}{R}\sum_{r=1}^{R}\text{mse}_i^{(r)}, \text{ and } \hat{\theta}_i = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i^{(r)}.$$

The $\text{mse}_i^{(r)}$ is the estimated MSE of $\hat{\theta}_i^{(r)}$ for the i^{th} area. They are computed using the formulas given in Sections 2 and 3.

Table 4.3 reports the average $\overline{\text{RRMSE}}$ and $\overline{\text{RRmse}}$ over the 30 small areas. When the sample size $n = 10,$ $\overline{\text{RRMSE}}$ is 4.98% for EBLUP and 5.49% for the pseudo-EBLUP respectively. As expected (You and Rao 2002), the pseudo-EBLUP has a slightly larger RRMSE than the one associated with EBLUP. Both the unit level EBLUP and pseudo-EBLUP estimators have much smaller RRMSE's than the area level estimators. For area level models, FH-HT and FH-HA perform similarly, with corresponding average true RRMSE equal to 9.72% and 9.68% respectively, when $n = 10.$ The FH-SRS performs poorly under informative sampling with the average true RRMSE equal to 18.89% when $n = 10.$ Even when $n = 30,$ the average RRMSE for FH-SRS is as large as 18.62%. Note that $\overline{\text{RRmse}}$ is very close to its true value.

In summary, the results in Table 4.3 show that the unit level estimators EBLUP and pseudo-EBLUP perform better than the area level estimators FH-HT and FH-HA under correct modeling. Both the area level estimators FH-HT and FH-HA perform reasonably well under informative sampling. As expected, FH-SRS performs poorly.

Table 4.3
Average RRMSE% for scenario I

		***n* = 10**		***n* = 30**	
Type	**Estimator**	$\overline{\textbf{RRMSE}}$	$\overline{\textbf{RRmse}}$	$\overline{\textbf{RRMSE}}$	$\overline{\textbf{RRmse}}$
Unit level	EBLUP	4.98	5.09	3.01	3.13
	Pseudo-EBLUP	5.49	5.66	3.58	3.67
Area level	FH-SRS	18.89	17.53	18.62	16.34
	FH-HT	9.72	10.25	6.67	6.69
	FH-HA	9.68	9.71	6.51	6.63

Table 4.4 displays the results of the average RRMSE under scenario II. The pseudo-EBLUP is the most robust estimator and has the smallest $\overline{\text{RRMSE}}$: the $\overline{\text{RRMSE}}$'s are 5.42% and 3.21% for $n = 10$ and $n = 30$ respectively. For the area level estimators, FH-HT and FH-HA perform similarly, whereas FH-SRS performs poorly. When $n = 10,$ $\overline{\text{RRMSE}}$ for FH-HT is 11.68% and 11.21% for FH-HA. When $n = 30,$ $\overline{\text{RRMSE}}$ is 7.24% for FH-HT and 6.79% for FH-HA. As expected, FH-SRS has large $\overline{\text{RRMSE}}$ under informative sampling. The pseudo-EBLUP performs the best in terms of bias, standard errors and RRMSE under model misspecification. FH-HA is slightly better than FH-HT. The estimated $\overline{\text{RRmse}}$ is very close to the true $\overline{\text{RRMSE}}$ for all estimators.

Table 4.4
Average RRMSE% for scenario II

Type	Estimator	$n = 10$		$n = 30$	
		$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$	$\overline{\text{RRMSE}}$	$\overline{\text{RRmse}}$
Unit level	EBLUP	6.78	6.94	5.62	5.81
	Pseudo-EBLUP	5.42	5.45	3.21	3.26
Area level	FH-SRS	19.76	17.43	19.06	16.24
	FH-HT	11.68	11.78	7.24	7.26
	FH-HA	11.21	11.27	6.79	6.91

4.2.3 Comparison of confidence intervals

We now compare the confidence intervals associated with the unit level and area level estimators. The confidence interval is in the form $\text{estimator} \pm z_{\alpha/2}\sqrt{\text{mse}}$, with $z_{\alpha/2}$ denoting the $100(1-\alpha/2)\%$ percentile of the standard normal distribution. For example, the 95% confidence interval of the EBLUP estimator $\hat{\theta}_i^{\text{EBLUP}}$ is obtained as $\hat{\theta}_i^{\text{EBLUP}} \pm 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right)}$, where $\text{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right)$ is given by (2.6). The confidence intervals are computed as follows. For a given estimator $\hat{\theta}_i^{(r)}$, $r = 1, \ldots, R$, $i = 1, \ldots, m$, define the indicator variable $I_i^{(r)}$ as:

$$I_i^{(r)} = \begin{cases} 1 \text{ if } & \theta_i \subseteq \left(\hat{\theta}_i^{(r)} - 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{(r)}\right)}, \hat{\theta}_i^{(r)} + 1.96\sqrt{\text{mse}\left(\hat{\theta}_i^{(r)}\right)}\right) \\ 0 \text{ otherwise} & \end{cases}.$$

The confidence interval coverage rate is obtained as the average of $I_i^{(r)}$ over the $R = 3{,}000$ simulations. Tables 4.5 and 4.6 present the 95% confidence interval coverage rates for the unit level and area level estimators under scenario I. The correlation coefficient ρ between the selection probabilities p_{ij} and y_{ij} is presented in the first column to reflect the strength of informativeness of the PPS sampling.

Table 4.5
Confidence interval coverage rates under scenario I: $n = 10$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.932	0.946	0.618	0.898	0.911
0.88	0.945	0.948	0.649	0.882	0.908
0.75	0.948	0.948	0.705	0.863	0.911
0.51	0.944	0.949	0.825	0.845	0.916
0.28	0.947	0.951	0.901	0.822	0.917
0.12	0.948	0.949	0.924	0.778	0.893
0.02	0.948	0.951	0.925	0.595	0.886
Average rate	*0.945*	*0.949*	*0.792*	*0.812*	*0.906*

We first discuss the coverage properties associated with the unit level estimators EBLUP and pseudo-EBLUP. These tables show that, when the model is correct, the coverage rates for EBLUP and pseudo-EBLUP are quite stable: the pseudo-EBLUP has slightly better coverage rate than EBLUP. When the sample size $n = 10,$ the average coverage rate for EBLUP is 94.5%, and 94.9% for pseudo-EBLUP. When the sample size $n = 30,$ it is 93.4% for EBLUP and 94.8% for pseudo-EBLUP. As the sample size increases from $n = 10$ to 30, the coverage rates for EBLUP deteriorate slightly more than those associated with the pseudo-EBLUP. The pseudo-EBLUP estimator is not as much affected by the degree of informativeness caused by the PPS sampling. The relatively stable coverage rates for EBLUP show that the sample is noninformative with respect to the correct unit level model. However, when $n = 30,$ EBLUP has slightly lower coverage rate.

Table 4.6
Confidence interval coverage rates under scenario I: $n = 30$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.905	0.946	0.265	0.932	0.926
0.88	0.938	0.948	0.286	0.915	0.921
0.75	0.941	0.949	0.377	0.911	0.924
0.51	0.940	0.951	0.625	0.895	0.931
0.28	0.941	0.950	0.806	0.874	0.929
0.12	0.939	0.945	0.923	0.866	0.922
0.02	0.937	0.948	0.937	0.772	0.917
Average rate	*0.934*	*0.948*	*0.603*	*0.881*	*0.924*

We now turn to the coverage rates associated with the area level estimators. As expected, FH-SRS has low coverage rates when the sampling is informative, and the coverage rate increases as the sampling design becomes non-informative. FH-HA has better coverage rate than FH-HT. The coverage rate for FH-HT decreases as the sampling design becomes non-informative. For example, when sample size $n = 10,$ the coverage rate for FH-HT is only 59.5% when the sampling is non-informative, compared to 88.6% of the coverage rate for FH-HA. As the sample size increases, the coverage rate for FH-HT and FH-HA improves. The average coverage rate for FH-HA is 90.6% when $n = 10$ and 92.4% when $n = 30.$ FH-HT has a lower coverage rate than the one associated with FH-HA. The average coverage rate is only 81.2% for FH-HT when $n = 10.$ The coverage rate for FH-SRS is very poor, 61.8%, under informative sampling when $n = 10$ and 26.5% when $n = 30.$ As the sample size increases, the coverage rate decreases for FH-SRS under informative sampling. As expected, the coverage rate gradually increases for FH-SRS as the sampling becomes non-informative. Among all the estimators, for both sample size $n = 10$ and $n = 30,$ the pseudo-EBLUP has the best coverage rate: FH-HA has the second best coverage rate.

Tables 4.7 and 4.8 present the coverage rates under scenario II. The results show that the EBLUP has low coverage rate under informative sampling, whereas the pseudo-EBLUP has very stable and high coverage rates (all around and over 95%) under both the informative and non-informative sampling. For example, when $n = 10,$ EBLUP has 84.6% coverage rate under informative sampling (correlation coefficient is 0.95), and when sample size increases to $n = 30,$ EBLUP has an even lower coverage rate of 62.9%. The average coverage rate is 90.4% for $n = 10$ and 79.6% for $n = 30$ for EBLUP under incorrect modeling. The results show that EBLUP is sensitive to the modeling when the sampling is informative. This is because EBLUP is completely model-based and ignores the sample design.

Table 4.7
Confidence interval coverage rates under scenario II: $n = 10$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.846	0.965	0.701	0.865	0.896
0.88	0.855	0.964	0.729	0.887	0.893
0.75	0.881	0.962	0.787	0.873	0.898
0.51	0.921	0.961	0.872	0.848	0.898
0.28	0.936	0.961	0.912	0.843	0.887
0.12	0.945	0.955	0.917	0.765	0.867
0.02	0.943	0.951	0.913	0.592	0.838
Average rate	*0.904*	*0.959*	*0.833*	*0.811*	*0.883*

Table 4.8
Confidence interval coverage rates under scenario II: $n = 30$

Correlation coefficient (ρ)	EBLUP	Pseudo-EBLUP	FH-SRS	FH-HT	FH-HA
0.95	0.629	0.969	0.239	0.913	0.923
0.88	0.638	0.965	0.275	0.895	0.919
0.75	0.708	0.964	0.406	0.908	0.923
0.51	0.829	0.963	0.701	0.923	0.926
0.28	0.902	0.964	0.854	0.911	0.921
0.12	0.931	0.958	0.921	0.884	0.912
0.02	0.937	0.953	0.918	0.778	0.894
Average rate	*0.796*	*0.962*	*0.616*	*0.887*	*0.918*

Among the three area level estimators, FH-HA performs the best. The coverage rate for FH-HA is very stable, and the average coverage rate for FH-HA is 88.3% when $n = 10$ and 91.8% when $n = 30.$ FH-HT has lower coverage rate when the sampling is very non-informative, particularly when sample size $n = 10.$ The average coverage rate for FH-HT is only 81.1% when $n = 10$ and 88.7% when $n = 30.$ The results show that FH-HA is superior to FH-HT. FH-SRS performs poorly when the sampling is informative, particularly when the sample size $n = 30.$ However, FH-SRS performs relatively well when the sampling becomes non-informative. The average coverage rate for FH-SRS is 83.3% when $n = 10,$ but only 61.6% when the sample size $n = 30.$

It is clear that pseudo-EBLUP has very high and stable coverage rate under incorrect modeling. FH-HA also has very stable but slightly lower coverage rate. Both EBLUP and FH-SRS have lower coverage rate as the sample size increases, especially when the sampling is informative.

5 Application to real data

In this section, we compare the unit level and area level estimates through a real data analysis. The data set we studied is the corn and soybean data provided by Battese et al. (1988). They considered the estimation of mean hectares of corn and soybeans per segment for twelve counties in north-central Iowa. Among the twelve counties, there were three counties with a single sample segment. We combined these three counties into a single one, resulting in 10 counties in our data set with sample size n_i ranging from 2 to 5 in each county. The total number of segments N_i (population size) within each county ranged from 402 to 1,505. Following You and Rao (2002), we assumed simple random sampling within each county, and the basic survey weight was computed as $w_{ij} = N_i / n_i$. For unit level modeling, y_{ij} is the number of hectares of corn (or soybean) in the j^{th} segment of the i^{th} county, the auxiliary variables are the number of pixels classified as corn and soybeans as in Battese et al. (1988). We applied the unit level model to the modified data set and obtained the EBLUP and pseudo-EBLUP estimates. For area level modeling, we first obtained the area level direct sample estimates $\hat{\theta}_i^{\text{SRS}}$ based on the SRS sampling. Next, we applied the Fay-Herriot model to the area level direct estimates and obtained the FH-SRS area level estimates. Figure 5.1 compares the area level direct estimates with the model-based unit level and area level estimates. In terms of point estimation, the EBLUP and pseudo-EBLUP estimates are almost identical as in You and Rao (2002). This is because the unit level model is a correct model for these data (Battese et al. 1988). The model-based area level estimates FH-SRS and the area level direct estimates are quite similar in this example.

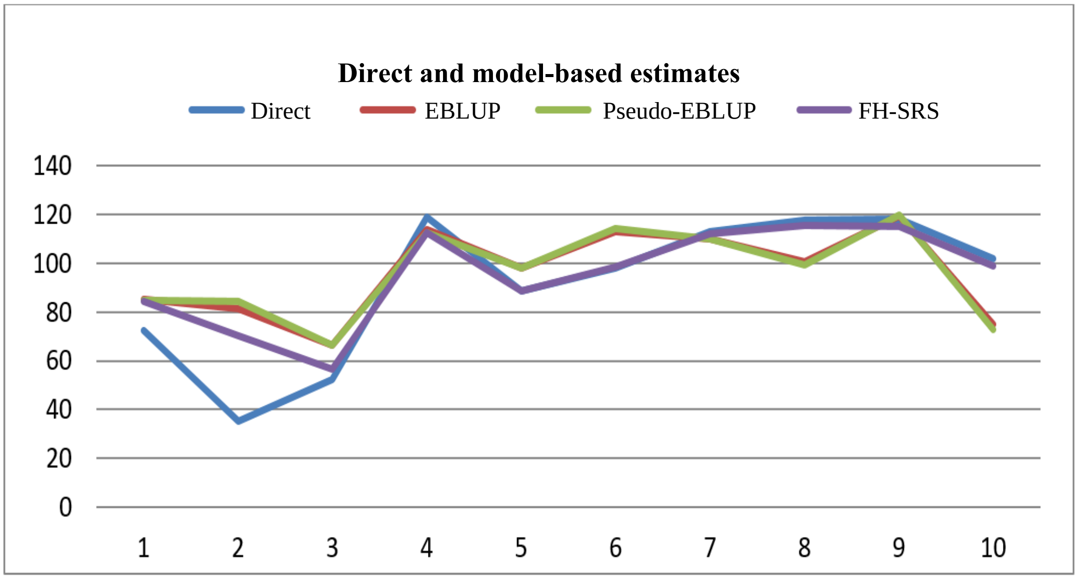

Figure 5.1 Comparison of direct and model-based estimates.

Figure 5.2 compares the standard errors of the direct and model-based estimators. The standard errors of the model-based estimators are the squared root of the estimated MSE. Both the unit level estimators EBLUP and pseudo-EBLUP have small and stable standard errors. As expected, pseudo-EBLUP has slightly larger standard errors than EBLUP. It is clear that the direct and FH-SRS standard errors are very variable and are very unstable. This example shows the effectiveness of the unit level EBLUP and pseudo-EBLUP estimators.

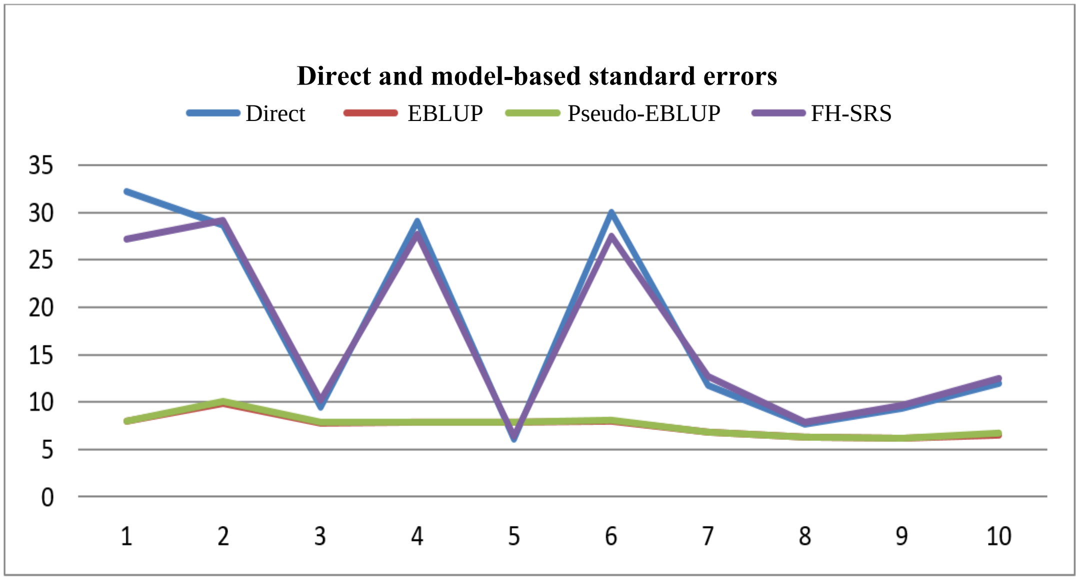

Figure 5.2 Comparison of direct and model-based standard errors.

6 Conclusions

In this paper, we compared 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. We compared the point estimates and coverage rate of confidence intervals of unit level and area level estimators. Overall, the unit level pseudo-EBLUP estimator performs the best in terms of bias and coverage rate under both informative and non-informative sampling. The EBLUP estimator performs well under correct modeling since the sampling is noninformative under correct unit level model (2.2). The pseudo-EBLUP estimator is also quite robust to the model misspecification as well. In practice, we suggest to construct the pseudo-EBLUP estimators using the survey weights and the unit level observations as discussed in Section 2.2. For area level models, FH-HA performs better than FH-HT, and FH-SRS performs poorly. We therefore recommend to construct the weighted HA estimators and then apply the Fay-Herriot model to obtain the corresponding model-based estimators if area level small area estimators are used.

Acknowledgements

The authors would like to thank an Associate Editor and two referees for their suggestions and comments to help to improve the presentation of the results substantially. In particular we would like to thank one referee’s very careful and constructive comments.

References

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.

Cressie, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 1, 75-94.

Datta, G.S., and 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.

Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

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*. New York: John Wiley & Sons, Inc.

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).

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

Verret, F., Rao, J.N.K. and Hidiroglou, M.A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41, 2, 333-347.

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. (2010). *Small Area Estimation under the Fay-Herriot Model Using Different Model Variance Estimation Methods and Different Input Sampling Variances*. Methodology branch working paper, SRID-2010-003E, Statistics Canada, Ottawa, Canada.

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.

You, Y., Rao, J.N.K. and Kovacevic, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. Proceedings: Symposium 2003, *Challenges in Survey Taking for the Next Decade*, Statistics Canada.