

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

## Survey Methodology

# Model-based small area estimation under informative sampling

by François Verret, J.N.K. Rao and Michael A. Hidioglou

Release date: December 17, 2015



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](http://www.statcan.gc.ca).

You can also contact us by

email at [STATCAN.infostats-infostats.STATCAN@canada.ca](mailto: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](http://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<sup>s</sup> value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
- <sup>P</sup> preliminary
- <sup>r</sup> revised
- X suppressed to meet the confidentiality requirements of the *Statistics Act*
- <sup>E</sup> 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

© Minister of Industry, 2015

All rights reserved. Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

**An HTML version is also available.**

*Cette publication est aussi disponible en français.*

---

# Model-based small area estimation under informative sampling

François Verret, J.N.K. Rao and Michael A. Hidirolou<sup>1</sup>

## Abstract

Unit level population models are often used in model-based small area estimation of totals and means, but the models may not hold for the sample if the sampling design is informative for the model. As a result, standard methods, assuming that the model holds for the sample, can lead to biased estimators. We study alternative methods that use a suitable function of the unit selection probability as an additional auxiliary variable in the sample model. We report the results of a simulation study on the bias and mean squared error (MSE) of the proposed estimators of small area means and on the relative bias of the associated MSE estimators, using informative sampling schemes to generate the samples. Alternative methods, based on modeling the conditional expectation of the design weight as a function of the model covariates and the response, are also included in the simulation study.

**Key Words:** Augmented model; Empirical best linear unbiased prediction (EBLUP); Nested error model; Pseudo-EBLUP.

## 1 Introduction

Estimates of population totals and means are often required for small subpopulations (or areas). Traditional area-specific direct estimators are not reliable if the area sample size is small. As a result, it becomes necessary to “borrow strength” across areas through indirect estimation based on models that provide a link to related areas. Linking models make use of auxiliary population information either at the area level or at the unit level. Rao (2003, Chapter 7) gives a detailed account of area level and unit level models that are widely used for small area estimation.

Suppose that the population of interest,  $U$ , consists of  $M$  non-overlapping areas with  $N_i$  elements in the  $i^{\text{th}}$  area ( $i = 1, \dots, M$ ). A sample,  $s$ , of  $m$  areas is first selected using a specified sampling scheme with inclusion probabilities  $\pi_i = mp_i$  ( $i = 1, \dots, M$ ), where  $p_i$  denotes the selection probability of area  $i$ . Subsamples  $s_i$  of specified sizes  $n_i$  are then independently selected from the sampled areas  $i$  according to specified sampling schemes with selection probabilities  $p_{j|i}$  ( $\sum_{j=1}^{N_i} p_{j|i} = 1$ ) such that the second-stage inclusion probabilities are  $\pi_{j|i} = n_i p_{j|i}$  for unit  $j$  in area  $i$  ( $j = 1, \dots, N_i$ ). Typically, the selection probability  $p_{j|i} = b_{ij} / \sum_{k=1}^{N_i} b_{ik}$ , where  $b_{ij}$  is a size measure related to the response variable  $y_{ij}$ . In this paper, we focus on the special case where all the areas are sampled,  $m = M$ .

We assume a nested error linear regression model for the population, based on covariates  $\mathbf{x}_{ij}$  related to the response variable  $y_{ij}$ . The population model is assumed to be given by

---

1. François Verret, Statistics Canada, 23 B, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: francois.verret@canada.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: jr Rao@math.carleton.ca; Michael A. Hidirolou, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: mike.hidirolou@canada.ca.

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}; j = 1, \dots, N_i; i = 1, \dots, M, \quad (1.1)$$

where  $v_i \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2)$  are random small area effects that are independent of the unit-level errors  $e_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma_e^2)$ ,  $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, \dots, x_{ijp})^T$  and  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ . Parameters of interest are the small area means  $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$  which may be approximated by  $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta} + v_i$ , if the area sizes  $N_i$  are large, where  $\bar{\mathbf{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$  is the known population mean of  $\mathbf{x}$  for area  $i$ .

Efficient model-based estimators of the area means  $\mu_i$  may be obtained if the sampling design is non-informative for the model, which implies that the sample and the population models coincide. In particular, empirical best linear unbiased prediction (EBLUP) estimators (Henderson 1975), based on the assumed sample model under non-informative sampling, may be used to estimate small area means  $\bar{Y}_i$  or  $\mu_i$  (see Section 2 and Rao 2003, Chapter 7). However, in many practical situations the selection probabilities  $p_{j|i}$  may be related to the associated  $y_{ij}$  even after conditioning on the covariates  $\mathbf{x}_{ij}$ . In such cases, we have “informative sampling” in the sense that the population model (1.1) no longer holds for the sample. For example, Pfeffermann and Sverchkov (2007) assumed that the sampled unit design weight  $w_{ji} = \pi_{ji}^{-1}$  is random with conditional expectation

$$\begin{aligned} E_{s_i}(w_{ji} | \mathbf{x}_{ij}, y_{ij}, v_i) &= E_{s_i}(w_{ji} | \mathbf{x}_{ij}, y_{ij}) \\ &= k_i \exp(\mathbf{x}_{ij}^T \mathbf{a} + by_{ij}), \end{aligned} \quad (1.2)$$

where  $\mathbf{a}$  and  $b$  are fixed unknown constants and

$$k_i = N_i n_i^{-1} \left\{ \sum_{j=1}^{N_i} \exp(-\mathbf{x}_{ij}^T \mathbf{a} - by_{ij}) / N_i \right\}.$$

Under informative sampling within areas, the EBLUP estimator of  $\bar{Y}_i$ , assuming that model (1.1) holds for the sample, may be heavily biased. It is, therefore, necessary to develop estimators that can account for sample selection bias and thus reduce estimation bias. Pfeffermann and Sverchkov (2007) developed a bias-adjusted estimator of the mean  $\bar{Y}_i$  under the assumption (1.2) on the design weights  $w_{ji}$  and assuming that the sample model is a nested error model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\alpha} + u_i + h_{ij}; j = 1, \dots, n_i; i = 1, \dots, M, \quad (1.3)$$

where  $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2)$ , and  $h_{ij} | j \in s_i \stackrel{\text{iid}}{\sim} N(0, \sigma_h^2)$ . Pfeffermann and Sverchkov (2007) noted that under a sampling scheme satisfying (1.2) the population model is also a nested error model but with different parameters. However, they do not use the form of the population model. The sample model (1.3) is identified after fitting the model to the sample data and then doing some model diagnostics. Similarly, model (1.2) on the weights is identified from the sample data  $\{w_{ji}, y_{ij}, \mathbf{x}_{ij}, j \in s_i, i \in S\}$ . Their estimators are noted (PS) in the following.

Prasad and Rao (1999) and You and Rao (2002) developed pseudo-EBLUP estimators of small area means  $\mu_i$  that depend on the sampling weights  $w_{ji}$ , assuming non-informative sampling for the model

(1.1). Their motivation for pseudo-EBLUP is to ensure design consistency as the area sample size,  $n_i$ , increases. The estimators of You and Rao (note (YR) in the following) also satisfy a benchmarking property in the sense that the associated estimators of area totals add up to a reliable direct estimator of the total, unlike the EBLUP estimators. Stefan (2005) studied the empirical performance of pseudo-EBLUP estimators under informative sampling for model (1.1) and showed that the pseudo-EBLUP leads to smaller bias compared to the EBLUP.

The main purpose of our paper is to study augmented sample models of the form

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + g(p_{j|i})\delta_0 + \tilde{v}_i + \tilde{e}_{ij}; j = 1, \dots, n_i; i = 1, \dots, M \tag{1.4}$$

for a suitably defined function  $g_{ij} = g(p_{j|i})$  of the probability  $p_{j|i}$ , where  $\tilde{v}_i \stackrel{iid}{\sim} N(0, \sigma_{v0}^2)$  and independent of  $\tilde{e}_{ij} \stackrel{iid}{\sim} N(0, \sigma_{e0}^2)$ , and  $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \dots, \beta_{0p})^T$ . The sample model (1.4) is identified after fitting the model to sample data for different choices of the function  $g(\cdot)$  and checking their adequacy. For example, residuals  $r_{ij}$  from fitting the model (1.4) without the augmenting variable  $g(p_{j|i})$  may be plotted against  $g(p_{j|i})$  to select  $g(\cdot)$ . The identified augmented sample model will also hold for the population (Skinner 1994, Rao 2003, Section 5.3). Possible choices of  $g(p_{j|i})$  are  $p_{j|i}$ ,  $\log p_{j|i}$ ,  $w_{j|i}$  and  $n_i w_{j|i} = p_{j|i}^{-1}$ .

From the augmented sample model (1.4) we obtain the EBLUP estimators of  $\bar{Y}_i$  or  $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta}_0 + \bar{G}_i \delta_0 + \tilde{v}_i$ , the approximate area mean under the augmented population model, where  $\bar{G}_i$  is the area mean of the population values  $g(p_{j|i}) \equiv g_{ij}$ . The EBLUP of  $\bar{Y}_i$  or  $\mu_i$  requires the knowledge of  $\bar{G}_i$  which depends on all the population values  $p_{j|i}$ . However, the choice  $g(p_{j|i}) = p_{j|i}$  gives  $\bar{G}_i = 1/N_i$  and the choice  $g(p_{j|i}) = n_i w_{j|i}$  gives  $\bar{G}_i = n_i \bar{W}_i$ , where  $\bar{W}_i$  is the area population mean of the weights  $w_{j|i}$ . The means  $\bar{W}_i$  are often known in practice. Pseudo-EBLUP estimators under the augmented model are also studied.

We conducted a simulation study under the design-model (or pm) framework to study the bias and MSE of the proposed estimators relative to EBLUP and pseudo-EBLUP estimators based on non-informative sampling, and the bias-adjusted estimators of Pfeffermann and Sverchkov (2007). We also studied the performance of MSE estimators in terms of relative bias.

Section 2 summarizes the existing model-based methods for estimating the small area means  $\bar{Y}_i$  or  $\mu_i$ . Proposed methods based on the augmented sample model (1.4) are presented in Section 3. The results of the simulation study are reported in Section 4. Concluding remarks are given in Section 5.

## 2 Existing methods

### 2.1 Estimators of small area means

Suppose that the population model (1.1) holds for the sample. Then the EBLUP estimator of  $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta} + v_i$  is given by

$$\hat{\mu}_i^H = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} + \hat{v}_i = \hat{\gamma}_i \bar{y}_i + (\bar{\mathbf{X}}_i - \hat{\gamma}_i \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}}, \tag{2.1}$$

where  $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i)$ ,  $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ ,  $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} / n_i$  are the unweighted sample means of the response variable  $y$  and the covariates  $\mathbf{x}$  and  $\hat{v}_i = \hat{\gamma}_i (\bar{y}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}})$ . Further,

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_i)^T \right\}^{-1} \left\{ \sum_{i=1}^M \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_i) y_{ij} \right\} \tag{2.2}$$

and  $\hat{\sigma}_e^2$  and  $\hat{\sigma}_v^2$  are obtained by the method of fitting of constants (Battese, Harter and Fuller (1988); Rao (2003, Chapter 7)) or restricted maximum likelihood (REML). The EBLUP estimator of the area mean  $\bar{Y}_i$  may be written in terms of  $\hat{\mu}_i^H$  as

$$\hat{Y}_i^H = N_i^{-1} \left[ (N_i - n_i) \hat{\mu}_i^H + n_i \left\{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}} \right\} \right], \tag{2.3}$$

(see Rao 2003, page 141). Note that  $\hat{Y}_i^H \approx \hat{\mu}_i^H$  if the sampling fraction  $n_i / N_i$  is sufficiently small. The EBLUP estimator  $\hat{Y}_i^H$  is design consistent under simple random sampling (SRS) or stratified SRS with proportional allocation within area  $i$ , leading to equal  $\pi_{ji}$ .

The pseudo-EBLUP estimator of  $\mu_i$  is given by

$$\hat{\mu}_i^{YR} = \hat{\gamma}_{iw} \bar{y}_{iw} + (\bar{\mathbf{X}}_i - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw})^T \hat{\boldsymbol{\beta}}_w, \tag{2.4}$$

where we denote by  $\tilde{w}_{j|i} = w_{ji} / \sum_{k=1}^{n_i} w_{k|i}$  the normalized weights,  $\hat{\gamma}_{iw} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \delta_i^2 \hat{\sigma}_e^2)$  with  $\delta_i^2 = \sum_{j=1}^{n_i} \tilde{w}_{j|i}^2$ ,  $\bar{y}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{j|i} y_{ij}$ ,  $\bar{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{j|i} \mathbf{x}_{ij}$  are the  $i^{\text{th}}$  area weighted means of  $y$  and  $\mathbf{x}$ , and

$$\hat{\boldsymbol{\beta}}_w = \left[ \sum_{i=1}^M \sum_{j=1}^{n_i} w_{ji} \mathbf{x}_{ij} (\mathbf{x}_{ij} - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw})^T \right]^{-1} \left[ \sum_{i=1}^M \sum_{j=1}^{n_i} w_{ji} (\mathbf{x}_{ij} - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw}) y_{ij} \right]. \tag{2.5}$$

The pseudo-EBLUP estimator  $\hat{\mu}_i^{YR}$  is design consistent under arbitrary selection probabilities  $p_{ji}$  unlike the EBLUP  $\hat{Y}_i^H$ .

Pfeffermann and Sverchkov (2007) studied estimation of small area means under informative sampling, assuming model (1.3) for the sample data and model (1.2) for the weights  $w_{ji}$ . Under this assumption, they obtained an estimator of  $\bar{Y}_i$  that provides protection against informative sampling. It is given by

$$\hat{Y}_i^{PS} = N_i^{-1} \left[ (N_i - n_i) \hat{\mu}_{iu}^H + n_i \left\{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\alpha}} \right\} + (N_i - n_i) \hat{b} \hat{\sigma}_h^2 \right], \tag{2.6}$$

where  $\hat{\mu}_{iu}^H = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\alpha}} + \hat{u}_i$  is the EBLUP estimator of  $\mu_{iu} = \bar{\mathbf{X}}_i^T \boldsymbol{\alpha} + u_i$  under the sample model (1.3) and  $\hat{b}$  is an estimator of  $b$  in the model (1.2) for the weights  $w_{ji}$ . Note that  $(\hat{\boldsymbol{\alpha}}, \hat{u}_i, \hat{\sigma}_u^2, \hat{\sigma}_h^2)$  is identical to  $(\hat{\boldsymbol{\beta}}, \hat{v}_i, \hat{\sigma}_v^2, \hat{\sigma}_e^2)$  obtained by assuming that the population model (1.1) holds for the sample. Therefore, we can also express  $\hat{Y}_i^{PS}$  as

$$\hat{Y}_i^{PS} = \hat{Y}_i^H + \left(1 - \frac{n_i}{N_i}\right) \hat{b} \hat{\sigma}_e^2, \tag{2.7}$$

noting that  $\hat{\mu}_i^H = \hat{\mu}_{iw}^H$ .

The last term in (2.7) corrects for any bias due to informative sampling under (1.2). PS obtained the estimator  $\hat{b}$  of  $b$  in (2.6) by regressing the sampling weights  $w_{ji}$  on  $k_i \exp(\mathbf{x}_{ij}^T \mathbf{a} + by_{ij})$ . The coefficients  $k_i$ ,  $\mathbf{a}$  and  $b$  may be estimated by fitting the model (1.2) using procedure NLIN in SAS or function nls in Splus. This involves iterative calculations and the initial values for  $\mathbf{a}$  and  $b$  are obtained by regressing  $\log(w_{ji})$  on  $\mathbf{x}_{ij}$  and  $y_{ij}$ . Initial values for  $k_i, i = 1, \dots, M$  are taken as  $k_i = N_i/n_i$ .

### 2.2 MSE estimation

The mean squared error (MSE) of the EBLUP estimator  $\hat{\mu}_i^H$ , assuming non-informative sampling, is estimated by

$$\text{mse}(\hat{\mu}_i^H) = g_{1i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2), \tag{2.8}$$

where

$$\begin{aligned} g_{1i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= (1 - \hat{\gamma}_i) \hat{\sigma}_v^2, g_{2i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) = (\bar{\mathbf{X}}_i - \hat{\gamma}_i \bar{\mathbf{x}}_i)^T \left( \sum_{i=1}^M \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} (\bar{\mathbf{X}}_i - \hat{\gamma}_i \bar{\mathbf{x}}_i), \\ g_{3i}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= \hat{\gamma}_i (1 - \hat{\gamma}_i)^2 \hat{\sigma}_e^{-4} \hat{\sigma}_v^{-2} h(\hat{\sigma}_e^2, \hat{\sigma}_v^2), \\ h(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= \hat{\sigma}_e^4 \text{var}(\hat{\sigma}_v^2) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \text{cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + \hat{\sigma}_v^4 \text{var}(\hat{\sigma}_e^2), \end{aligned}$$

$\hat{\mathbf{V}}_i = \hat{\sigma}_e^2 \mathbf{1}_{n_i} + \hat{\sigma}_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T$  and  $\mathbf{X}_i^T = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})$ . The matrix  $\sum_{i=1}^M \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i$  may be expressed explicitly as  $\hat{\sigma}_e^{-2} \sum_{i=1}^M \sum_{j=1}^{n_i} (\mathbf{x}_{ij} \mathbf{x}_{ij}^T - \hat{\gamma}_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T)$ . The MSE estimator (2.8) is unbiased to second order under non-informative sampling (Rao 2003, Chapter 7). We refer the reader to Rao (2003, page 142) for the corresponding MSE estimator of  $\hat{Y}_i^H$ .

The MSE of the pseudo-EBLUP estimator  $\hat{\mu}_i^{YR}$  is estimated by

$$\text{mse}(\hat{\mu}_i^{YR}) = g_{1iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + g_{2iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) + 2g_{3iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2), \tag{2.9}$$

where

$$\begin{aligned} g_{1iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= (1 - \hat{\gamma}_{iw}) \hat{\sigma}_v^2, g_{2iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) = (\bar{\mathbf{X}}_i - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw})^T \Phi_w (\bar{\mathbf{X}}_i - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw}), \\ \Phi_w &= \hat{\sigma}_e^2 \left( \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{z}_{ij}^T \right)^{-1} \left( \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}_{ij}^T \right) \left\{ \left( \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{z}_{ij}^T \right)^{-1} \right\}^T \\ &+ \hat{\sigma}_v^2 \left( \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{z}_{ij}^T \right)^{-1} \left\{ \sum_{i=1}^M \left( \sum_{j=1}^{n_i} \mathbf{z}_{ij} \right) \left( \sum_{j=1}^{n_i} \mathbf{z}_{ij} \right)^T \right\} \left\{ \left( \sum_{i=1}^M \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{z}_{ij}^T \right)^{-1} \right\}^T, \\ g_{3iw}(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= \hat{\gamma}_{iw} (1 - \hat{\gamma}_{iw})^2 \hat{\sigma}_e^{-4} \hat{\sigma}_v^{-2} h(\hat{\sigma}_e^2, \hat{\sigma}_v^2) \end{aligned}$$

and  $\mathbf{z}_{ij} = w_{ij} (\mathbf{x}_{ij} - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw})$ ; see You and Rao (2002). The MSE estimator (2.9) is obtained by ignoring a cross-product term in  $\text{MSE}(\hat{\mu}_i^{\text{YR}})$ . Torabi and Rao (2010) obtained a MSE estimator that accounts for the missing cross-product term and that is unbiased to second order under non-informative sampling. However, it is computationally more intensive than (2.9). It was not used in the simulation study (Section 4) since it would have slowed down the simulations significantly. A few simulation trials, however, revealed that the two MSE estimators give similar results under the simulation set-up used in Section 4.

Pfeffermann and Sverchkov (2007) proposed a parametric bootstrap method to estimate the MSE of the bias-adjusted estimator  $\hat{Y}_i^{\text{PS}}$  given by (2.6). We have not included this MSE estimator in our simulation study.

### 3 Proposed method

The proposed method of estimating the small area means,  $\bar{Y}_i$ , is simple. It uses the standard EBLUP estimator under the augmented sample model (1.4). The model parameters  $(\sigma_{v_0}^2, \sigma_{e_0}^2)$  and  $(\boldsymbol{\beta}_0, \delta_0)$  are estimated by REML and weighted least squares (WLS) respectively. The EBLUP estimator of  $\mu_i$  under the augmented model (1.4) is given by

$$\hat{\mu}_{i(a)}^H = \hat{\gamma}_{i0} \bar{y}_i + (\bar{\mathbf{X}}_i - \hat{\gamma}_{i0} \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}}_0 + (\bar{G}_i - \hat{\gamma}_{i0} \bar{g}_i) \hat{\delta}_0, \tag{3.1}$$

where  $\hat{\gamma}_{i0} = \hat{\sigma}_{v_0}^2 / (\hat{\sigma}_{v_0}^2 + \hat{\sigma}_{e_0}^2 / n_i)$ ,  $(\hat{\boldsymbol{\beta}}_0^T, \hat{\delta}_0)$  is the WLS estimator of  $(\boldsymbol{\beta}_0^T, \delta_0)$  and  $\bar{g}_i = \sum_{j=1}^{n_i} g_{ij} / n_i$ . Note that  $\hat{\mu}_{i(a)}^H$  assumes that  $\bar{G}_i$  is known. The EBLUP estimator of  $\bar{Y}_i$  under the augmented model may be written in terms of  $\hat{\mu}_{i(a)}^H$  as

$$\hat{Y}_{i(a)}^H = N_i^{-1} [(N_i - n_i) \hat{\mu}_{i(a)}^H + n_i \{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}}_0 + (\bar{G}_i - \bar{g}_i) \hat{\delta}_0 \}]. \tag{3.2}$$

The pseudo-EBLUP estimator of  $\mu_i$  under the augmented model (1.4) is similarly obtained by modifying (3.1) as

$$\hat{\mu}_{i(a)}^{\text{YR}} = \hat{\gamma}_{i0w} \bar{y}_{iw} + (\bar{\mathbf{X}}_i - \hat{\gamma}_{i0w} \bar{\mathbf{x}}_{iw})^T \hat{\boldsymbol{\beta}}_{0w} + (\bar{G}_i - \hat{\gamma}_{i0w} \bar{g}_{iw}) \hat{\delta}_{0w}, \tag{3.3}$$

where  $\hat{\gamma}_{i0w} = \hat{\sigma}_{v_0}^2 / (\hat{\sigma}_{v_0}^2 + \delta_i^2 \hat{\sigma}_{e_0}^2)$ ,  $\bar{g}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{ji} g_{ij}$  and  $(\hat{\boldsymbol{\beta}}_{0w}, \hat{\delta}_{0w})$  are obtained by suitably modifying (2.5).

The MSE estimators of  $\hat{\mu}_{i(a)}^H$  and  $\hat{\mu}_{i(a)}^{\text{YR}}$  under the augmented model (1.4) are obtained by suitably modifying (2.8) and (2.9) respectively. Note that we only need to apply existing formulae to the augmented sample model (1.4) to get the EBLUP and the pseudo-EBLUP estimators and associated MSE estimators. New software development is not needed.

Our main interest is to study the performance of the estimators of  $\bar{Y}_i$  based on the sample augmented model under informative sampling. Since the estimators  $\hat{Y}_{i(a)}^H$  and  $\hat{\mu}_{i(a)}^{\text{YR}}$  are obtained under the augmented

model (1.4), they are likely to perform well for the following reasons: (a) If the augmented model holds for the sample, then it also holds for the population, and the non-sampled values  $y_{ij}$  can be predicted by fitting the augmented model to the sample; (b) If the augmenting variable  $g_{ij}$  explains  $y_{ij}$  after conditioning on  $\mathbf{x}_{ij}$ , then  $\sigma_{e0}^2$  and  $\sigma_{v0}^2$  may be smaller than the corresponding  $\sigma_e^2$  and  $\sigma_v^2$  for the original population model, thus leading to better predictors of the non-sampled  $y_{ij}$ . Pfeffermann and Sverchkov (2003) demonstrated, under a different model setup, that the inclusion of sample selection probabilities in the model “can reduce the RMSE quite substantially”.

## 4 Simulation study

### 4.1 Implementation

A design-model (pm) approach was used for the simulation study by generating data for the  $N = \sum_{i=1}^M N_i$  population units according to a specified model, and then selecting a sample according to a specified design. The process of generating population data and then selecting a sample is repeated  $R$  times. We next describe the steps to implement the process. The population data,  $y_{ij}$ , for  $M = 99$  areas and  $N_i = 100$  units within each area  $i$  were generated from the simple nested error linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + e_{ij}; \quad i = 1, \dots, 99; \quad j = 1, \dots, 100, \quad (4.1)$$

where  $\beta_0 = 1, \beta_1 = 1, v_i \sim N(0, \sigma_v^2 = 0.5)$  and independent of  $e_{ij} \sim N(0, \sigma_e^2 = 2)$ . The population  $x_{ij}$  – values were generated from a gamma distribution with mean 10 and variance 50, and held fixed over the simulation of population  $y_{ij}$  – values from (4.1).

We considered different sample sizes within areas by fixing  $n_i = 5$  for the first 33 areas,  $n_i = 7$  for the next 33 areas and  $n_i = 9$  for the last 33 areas. This was done to study the effect of unequal sample sizes on the choice of the augmenting variable  $g_{ij} = g(p_{j|i})$ . Samples of specified sizes,  $n_i$ , were selected within the areas with probabilities proportional to specified sizes,  $b_{ij}$ , using the Rao-Sampford (Rao 1965 and Sampford 1967) method of sampling with unequal probabilities and without replacement. The latter method ensures that the inclusion probabilities  $\pi_{j|i}$  are proportional to the sizes  $b_{ij}$ , i.e.,  $\pi_{j|i} = n_i b_{ij} / B_i = n_i p_{j|i}$ ,  $j = 1, \dots, N_i$ , where  $B_i$  is the total of the  $b_{ij}$  in area  $i$ .

We considered two different choices of the sizes  $b_{ij}$  in the simulation study. The first choice uses

$$\begin{aligned} b_{ij} &= \exp\left[\left\{-\left(y_{ij} - \beta_0 - \beta_1 x_{ij}\right) / \sigma_e + \delta_{ij} / 5\right\} / 3\right] \\ &= \exp\left[\left\{-\left(v_i + e_{ij}\right) / \sigma_e + \delta_{ij} / 5\right\} / 3\right], \end{aligned} \quad (4.2)$$

where  $\delta_{ij} \stackrel{\text{iid}}{\sim} N(0, 1)$ . The size measures (4.2) are equivalent to those used by Pfeffermann and Sverchkov (2007) in their simulation study and satisfy the relationship (1.2) on the weights  $w_{ji} = \pi_{ji}^{-1}$ . Following PS, the area effects  $v_i$  and the unit errors  $e_{ij}$  were truncated to  $\pm 2.5\sigma_v$  and  $\pm 2.5\sigma_e$  to avoid extreme selection probabilities.

The second choice of size measures, following Asparouhov (2006), involves two different types of size measures: invariant (I) and non-invariant (NI). For the invariant case,  $b_{ij}$  is independent of  $v_i$  given  $\mathbf{x}_{ij}$ ; otherwise, it is called non-invariant. Invariant size measures are given by

$$b_{ij} = \left[ 1 + \exp \left\{ -\tau \left( \frac{1}{\alpha} e_{ij} + \sqrt{1 - \frac{1}{\alpha^2}} e_{ij}^* \right) \right\} \right]^{-1}. \quad (4.3)$$

Non-invariant size measures are taken as

$$b_{ij} = \left[ 1 + \exp \left\{ -\tau \left( \frac{1}{\alpha} (v_i + e_{ij}) + \sqrt{1 - \frac{1}{\alpha^2}} (v_i^* + e_{ij}^*) \right) \right\} \right]^{-1}. \quad (4.4)$$

The coefficient  $\tau$  in (4.3) and (4.4), chosen as 0.5, ensures that the variation of the weights  $w_{ji}$  would not be too large within a simulation run. The random pair  $(v_i^*, e_{ij}^*)$  was generated independently of  $(v_i, e_{ij})$  from the same distributions as  $v_i$  and  $e_{ij}$  to ensure that the weight variation would be comparable between various levels of  $\alpha$ . If some of the  $\pi_{ji}$  exceeded one, they were set to one, and the probabilities were recomputed for the remaining units. The  $\alpha$ -values in (4.3) and (4.4), chosen as 1, 2, 3 or  $\infty$ , control the level of informativeness. Increasing  $\alpha$  decreases informativeness, with  $\alpha = \infty$  corresponding to non-informative sampling. Various dependencies in the simulations were introduced as follows, in order to increase the precision of comparisons between different estimators: All the four error components  $(v_i, e_{ij}, v_i^*, e_{ij}^*)$  were first generated. Population  $y$ -values, as well as invariant and non-invariant probabilities of selection, were then generated from those errors. For a given generated population, eight samples were selected: an invariant sample and a non-invariant sample for each value of  $\alpha$  considered.

It may be noted that the weights  $w_{ji}$  obtained from the size measures (4.3) and (4.4) may not satisfy condition (1.2) of PS. We nevertheless fitted (1.2) to those weights to compute  $\hat{b}$  needed in the bias-adjusted estimator  $\hat{Y}_i^{\text{PS}}$ .

Using the design-model (pm) approach,  $R = 1,000$  samples were generated under the size measures (4.2) and the size measures (4.3) and (4.4). From each simulated sample  $r$  ( $r = 1, \dots, R$ ), the estimates  $\hat{Y}_i^{H(r)}$ ,  $\hat{Y}_{i(a)}^{H(r)}$  and  $\hat{Y}_i^{\text{PS}(r)}$  were computed for each small area  $i$ ; for the YR method only  $\hat{\mu}_i^{\text{YR}(r)}$  and  $\hat{\mu}_{i(a)}^{\text{YR}(r)}$  were computed. Also, the MSE estimates,  $\text{mse}(\hat{\mu}_i^H)^{(r)}$ ,  $\text{mse}(\hat{\mu}_{i(a)}^H)^{(r)}$ ,  $\text{mse}(\hat{\mu}_i^{\text{YR}})^{(r)}$  and  $\text{mse}(\hat{\mu}_{i(a)}^{\text{YR}})^{(r)}$ , associated with  $\hat{\mu}_i^H$ ,  $\hat{\mu}_{i(a)}^H$ ,  $\hat{\mu}_i^{\text{YR}}$  and  $\hat{\mu}_{i(a)}^{\text{YR}}$ , were computed. As noted earlier, we did not include the bootstrap MSE estimator of  $\hat{Y}_i^{\text{PS}}$ , proposed by Pfeffermann and Sverchkov (2007), in the simulation study. Also, for simplicity, we did not include the MSE estimators of  $\hat{Y}_i^H$  and  $\hat{Y}_{i(a)}^H$  because the latter estimators performed similarly to  $\hat{\mu}_i^H$  and  $\hat{\mu}_{i(a)}^H$  in terms of MSE.

We considered the following performance measures for a given estimator, say of the small area mean  $\bar{Y}_i$ . Average absolute bias ( $\overline{AB}$ ) is measured by

$$\overline{AB} = \frac{1}{M} \sum_{i=1}^M AB_i$$

with

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

where  $\hat{Y}_i^{(r)}$  and  $\bar{Y}_i^{(r)}$  are the values of  $\hat{Y}_i$  and  $\bar{Y}_i$  for the  $r^{\text{th}}$  simulated sample and population. Efficiency of an estimator  $\hat{Y}_i$  is measured by the average root MSE

$$\overline{RMSE} = \frac{1}{M} \sum_{i=1}^M \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{Y}_i^{(r)} - \bar{Y}_i^{(r)})^2}$$

Turning to the performance of MSE estimators  $mse(\hat{\mu}_i^H)$ ,  $mse(\hat{\mu}_{i(a)}^H)$ ,  $mse(\hat{\mu}_i^{YR})$  and  $mse(\hat{\mu}_{i(a)}^{YR})$  in estimating MSEs, we first calculated reliable measures of MSEs by increasing  $R = 1,000$  to  $T = 10,000$  simulated samples. The MSE of an estimator  $\hat{\mu}_i$  is then calculated as

$$MSE(\hat{\mu}_i) = \frac{1}{T} \sum_{t=1}^T (\hat{\mu}_i^{(t)} - \bar{Y}_i^{(t)})^2,$$

where  $\hat{\mu}_i^{(t)}$  and  $\bar{Y}_i^{(t)}$  denote the values of  $\hat{\mu}_i$  and  $\bar{Y}_i$  for the  $t^{\text{th}}$  simulated sample and population. For MSE estimation, we retained the original  $R$  simulated samples and calculated the expected values  $E[mse(\hat{\mu}_i)] = R^{-1} \sum_{r=1}^R mse(\hat{\mu}_i^{(r)})$ , where  $mse(\hat{\mu}_i^{(r)})$  denotes the value of the MSE estimate for the  $r^{\text{th}}$  simulated sample. The average absolute relative bias ( $\overline{ARB}$ ) of a MSE estimator  $mse(\hat{\mu}_i)$  is then calculated as

$$\overline{ARB}[mse(\hat{\mu}_i)] = M^{-1} \sum_{i=1}^M \left| \frac{E[mse(\hat{\mu}_i)]}{MSE(\hat{\mu}_i)} - 1 \right|.$$

## 4.2 Results under the Pfeffermann and Sverchkov size measures

Table 4.1 reports the simulation results on the average absolute bias ( $\overline{AB}$ ) and the average root mean square error ( $\overline{RMSE}$ ) of the estimators  $\hat{Y}_i^H$ ,  $\hat{Y}_{i(a)}^H$ ,  $\hat{\mu}_i^{YR}$ ,  $\hat{\mu}_{i(a)}^{YR}$  and  $\hat{Y}_i^{PS}$  under the PS size measures (4.2). The average absolute RB ( $\overline{ARB}$ ) of the MSE estimators  $mse(\hat{\mu}_i^H)$ ,  $mse(\hat{\mu}_{i(a)}^H)$ ,  $mse(\hat{\mu}_i^{YR})$  and  $mse(\hat{\mu}_{i(a)}^{YR})$  are also reported. Four different choices of the augmenting variable  $g_{ij}$  were studied:  $p_{j|i}$ ,  $w_{j|i}$ ,  $n_i w_{j|i} = p_{j|i}^{-1}$  and  $\log p_{j|i}$ . Bootstrap estimator of  $MSE(\hat{Y}_i^{PS})$ , proposed by Pfeffermann and Sverchkov (2007), is not included in our study because the bootstrap simulation is very computer intensive.

Table 4.1 shows that the  $\overline{AB}$  of the EBLUP estimator  $\hat{Y}_i^H$  is large ( $= 0.456$ ) relative to the corresponding augmented model EBLUP,  $\hat{Y}_{i(a)}^H$ , for the four choices of  $g_{ij}$ . Also, the choice  $g_{ij} = w_{j|i}$  leads to larger  $\overline{AB}$  compared to the other three choices (0.131 compared to 0.042 or less). The customary pseudo-EBLUP,  $\hat{\mu}_i^{YR}$ , surprisingly performed well ( $\overline{AB} = 0.044$ ) even though it was obtained under the assumption of noninformative sampling. This good performance is perhaps due to the use of weights in  $\hat{\mu}_i^{YR}$ . Augmented pseudo-EBLUP,  $\hat{\mu}_{i(a)}^{YR}$ , leads to further reduction in  $\overline{AB}$ . The PS estimator,  $\hat{Y}_i^{PS}$ , performs well relative to  $\hat{Y}_{i(a)}^H$  :  $\overline{AB} = 0.033$ .

Turning to  $\overline{RMSE}$ , Table 4.1 shows that  $\hat{Y}_i^H$  has the largest value ( $= 0.617$ ) due to large  $\overline{AB}$ , followed by  $\hat{\mu}_i^{YR}$  and  $\hat{Y}_i^{PS}$  with values 0.442 and 0.416 respectively. On the other hand, the augmented model estimators performed significantly better relative to  $\hat{Y}_i^{PS}$  and  $\hat{\mu}_i^{YR}$ . For example, the choice  $g_{ij} = p_{j|i}$  gives  $\overline{RMSE} = 0.151$ . Among the four choices of  $g_{ij}$ , the choice  $w_{j|i}$  gives the largest  $\overline{RMSE}$  ( $= 0.242$ ). We also calculated  $\overline{AB}$  and  $\overline{RMSE}$  of the approximate EBLUP estimators  $\hat{\mu}_i^H$  and  $\hat{\mu}_{i(a)}^H$ . We found that the values are practically the same as the corresponding values for  $\hat{Y}_i^H$  and  $\hat{Y}_{i(a)}^H$ .

Finally, with respect to MSE estimation,  $mse(\hat{\mu}_i^H)$  exhibits largest  $\overline{ARB}$  : 53.1% compared to 3.8% for  $\hat{\mu}_i^{YR}$ , although  $\overline{RMSE}$  for  $\hat{\mu}_i^{YR}$  is larger compared to  $\hat{\mu}_i^H$  based on  $p_{j|i}$  or  $n_i w_{j|i}$ . The MSE estimators  $mse(\hat{\mu}_{i(a)}^H)$  and  $mse(\hat{\mu}_{i(a)}^{YR})$  lead to small  $\overline{ARB}$  ( $< 7\%$ ) except for the choice  $w_{j|i}$  which leads to  $\overline{ARB} = 62.6\%$  for  $\hat{\mu}_{i(a)}^H$  and  $\overline{ARB} = 39.6\%$  for  $\hat{\mu}_{i(a)}^{YR}$ .

**Table 4.1**  
Average absolute bias ( $\overline{AB}$ ), average RMSE ( $\overline{RMSE}$ ) of the estimators and percent average absolute RB ( $\overline{ARB}$ ) of the MSE estimators: PS size measures

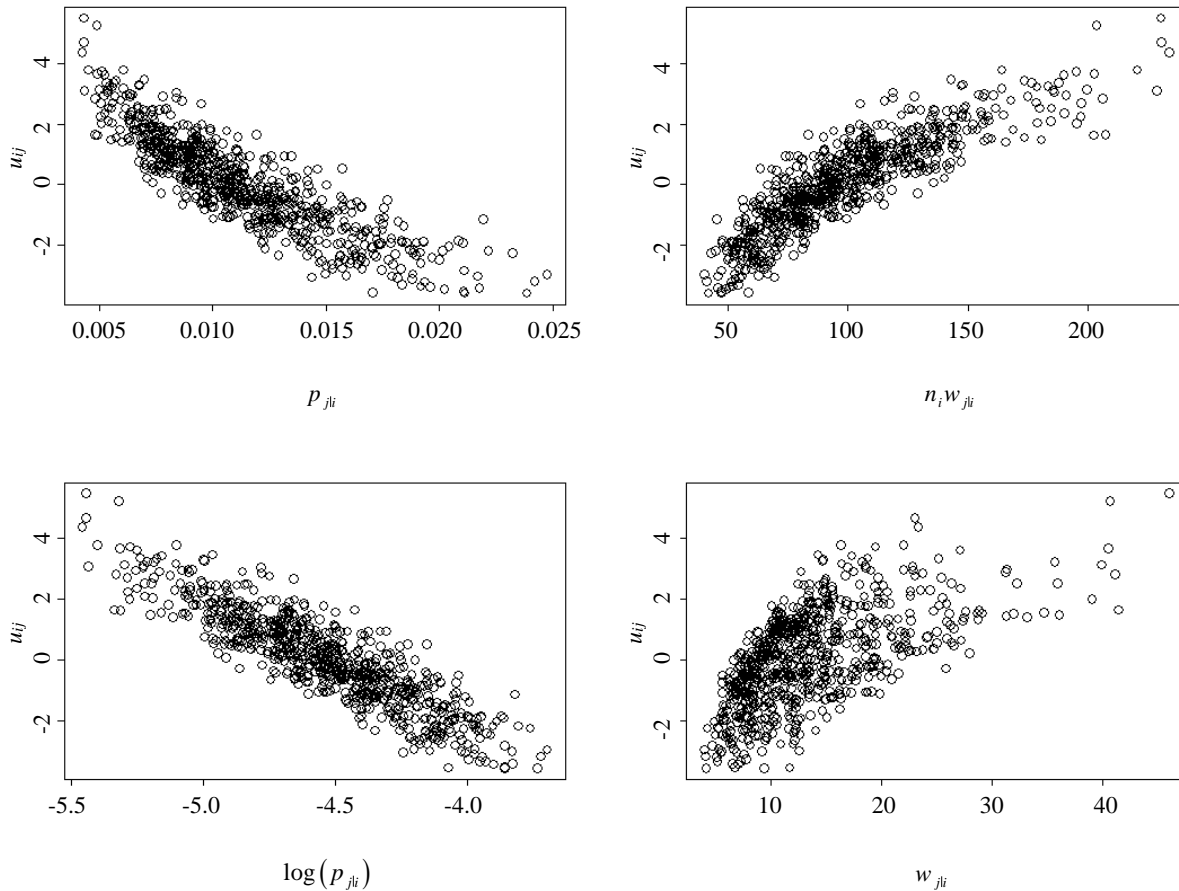
Performance measure	EBLUP					pseudo-EBLUP					PS
	$\hat{Y}_i^H$	$\hat{Y}_{i(a)}^H$				$\hat{\mu}_i^{YR}$	$\hat{\mu}_{i(a)}^{YR}$				$\hat{Y}_i^{PS}$
		$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$		$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$	
$\overline{AB}$	0.456	0.042	0.004	0.131	0.003	0.044	0.007	0.004	0.044	0.003	0.033
$\overline{RMSE}$	0.617	0.151	0.147	0.242	0.101	0.442	0.157	0.156	0.207	0.106	0.416
$\% \overline{ARB}(\text{mse})$	53.1	3.7	6.7	62.6	6.9	3.8	4.1	5.2	39.6	6.7	

### 4.3 Selection of the augmenting variable

In this section we illustrate the selection of the augmenting variable by generating data for the  $N$  population units from model (4.1) and then selecting a sample from the population data according to the Rao-Sampford method using size measures (4.2). Letting  $u_{ij} = v_i + e_{ij}$ , we fitted the model  $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$  to the sample data by ordinary least squares (OLS) and obtained the residuals  $\tilde{u}_{ij} = y_{ij} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{ij}$ , where  $\tilde{\beta}_0$  and  $\tilde{\beta}_1$  are the OLS estimators of  $\beta_0$  and  $\beta_1$  respectively.

Figure 4.1 gives residual plots of  $(\tilde{u}_{ij}, p_{j|i}), (\tilde{u}_{ij}, \log p_{j|i}), (\tilde{u}_{ij}, n_i w_{j|i})$  and  $(\tilde{u}_{ij}, w_{j|i})$ . All four plots clearly indicate informative sampling. Linear relationships between  $u_{ij}$  and the two choices  $p_{j|i}$  and

$\log p_{ji}$  suggest that either of them should work well. The choice  $w_{ji}$  indicates some non-linearity and wider scatter in the residual plot, and this choice led to the largest  $\overline{\text{RMSE}}$  among the four choices, as shown in Table 4.1. The choice  $n_i w_{ji}$  also indicates some non-linearity but less scatter in the residual plot.



**Figure 4.1 Residual plots for a simulated example: PS size measures.**

We have also fitted the augmented model (1.4) with  $g(p_{ji}) = p_{ji}$  and calculated the OLS residuals  $\tilde{u}_{0ij} = y_{ij} - \tilde{\beta}_{00} - \tilde{\beta}_{01}x_{ij} - \tilde{\delta}_0 p_{ji}$ . All the residuals  $\tilde{u}_{0ij}$  are less than 2.0 in absolute value, suggesting adequacy of the augmented model.

#### 4.4 Results under Asparouhov size measures

Table 4.2 reports the simulation results on  $\overline{\text{AB}}$  under the Asparouhov size measures (4.3) and (4.4). It shows, as in Table 4.1 for the PS size measures, that  $\overline{\text{AB}}$  of the EBLUP is large (0.437 for the invariant size measures (I) and 0.440 for non-invariant size measures (NI)) when the augmenting variable,  $g_{ij}$ , is

not included in the model and sampling is very informative ( $\alpha = 1$ ). Also,  $\overline{AB}$  decreases as  $\alpha$  increases. On the other hand, under the same model  $\overline{AB}$  associated with pseudo-EBLUP is much lower: 0.048 for I and 0.047 for NI when  $\alpha = 1$ , and  $\overline{AB}$  decreases as  $\alpha$  increases. The PS estimator under the same model also exhibits lower  $\overline{AB}$  (about 0.01) regardless of the choice of the value of  $\alpha$ . Inclusion of  $p_{j_i}$  or  $n_i w_{j_i}$  or  $\log p_{j_i}$  as augmenting variable in the model also leads to small  $\overline{AB}$  for the EBLUP (0.02 or less) regardless of the value of  $\alpha$ . On the other hand, the choice  $w_{j_i}$  as the augmenting variable leads to larger  $\overline{AB}$  (0.14 for  $\alpha = 1$  and 2), except for non-informative sampling ( $\alpha = \infty$ ). This poor performance of the choice  $w_{j_i}$  is probably due to the fact that  $w_{j_i} = (n_i p_{j_i})^{-1}$  depends on  $n_i$  when the area sample sizes,  $n_i$ , are not equal, unlike the other choices of  $g_{ij}$ . Pseudo-EBLUP performed similarly to EBLUP under the augmented model in terms of  $\overline{AB}$ .

**Table 4.2**  
Average absolute bias ( $\overline{AB}$ ) of the estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

$\alpha$	Size measure	EBLUP					pseudo-EBLUP				PS	
		$\hat{\bar{Y}}_i^H$	$\hat{\bar{Y}}_{i(a)}^H$				$\hat{\mu}_i^{YR}$	$\hat{\mu}_{i(a)}^{YR}$				$\hat{\bar{Y}}_i^{PS}$
			$p_{j_i}$	$n_i w_{j_i}$	$w_{j_i}$	$\log p_{j_i}$		$p_{j_i}$	$n_i w_{j_i}$	$w_{j_i}$	$\log p_{j_i}$	
1	I	0.437	0.001	0.005	0.140	0.022	0.048	0.001	0.006	0.057	0.005	0.012
	NI	0.440	0.007	0.007	0.145	0.021	0.047	0.003	0.007	0.064	0.005	0.013
2	I	0.217	0.009	0.010	0.137	0.014	0.024	0.010	0.010	0.098	0.010	0.012
	NI	0.217	0.011	0.009	0.136	0.011	0.024	0.009	0.010	0.098	0.010	0.012
3	I	0.145	0.010	0.010	0.101	0.011	0.017	0.010	0.010	0.075	0.010	0.011
	NI	0.144	0.011	0.011	0.099	0.012	0.016	0.010	0.011	0.074	0.011	0.011
$\infty$	I	0.011	0.011	0.011	0.011	0.011	0.012	0.011	0.011	0.012	0.011	0.011
	NI	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010

Table 4.3 reports the simulation results on the average root mean squared error ( $\overline{RMSE}$ ) using the Asparouhov size measures (4.3) and (4.4). It shows that the EBLUP, based on model (1.4) without the augmenting variable  $g_{ij}$ , has the largest  $\overline{RMSE}$  (0.596 for I and 0.619 for NI) when the sampling is very informative ( $\alpha = 1$ ). The  $\overline{RMSE}$  gradually decreases to around 0.42 as the sampling becomes non-informative ( $\alpha = \infty$ ). On the other hand,  $\overline{RMSE}$  of both the pseudo-EBLUP (without the  $g_{ij}$  – term in the model) and PS do not depend on  $\alpha$ , and lead to significant reduction:  $\overline{RMSE}$  of the pseudo-EBLUP is around 0.44 and  $\overline{RMSE}$  of PS is slightly smaller, around 0.42. Increase in  $\overline{RMSE}$  of the pseudo-EBLUP and PS over the EBLUP under non-informative sampling ( $\alpha = \infty$ ) is also small. On the other hand, EBLUP and pseudo-EBLUP under the augmented model lead to large reduction in MSE when the sampling is very informative ( $\alpha = 1$ ), particularly for the choices  $p_{j_i}$  and  $\log p_{j_i}$ :  $\overline{RMSE}$  less than 0.15. The choice of  $w_{j_i}$  leads to larger  $\overline{RMSE}$  (around 0.29) when  $\alpha = 1$  but it is still much smaller than the  $\overline{RMSE}$  for the pseudo-EBLUP without the  $g_{ij}$  – term and the PS. As  $\alpha$  increases,  $\overline{RMSE}$  is roughly the same for EBLUP (under the augmented model), pseudo-EBLUP and PS.

**Table 4.3**  
Average root mean squared error (RMSE) of the estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

$\alpha$	Size measure	EBLUP					pseudo-EBLUP					PS $\hat{Y}_i^{PS}$
		$\hat{Y}_i^H$	$\hat{Y}_{i(\alpha)}^H$				$\hat{\mu}_i^{VR}$	$\hat{\mu}_{i(\alpha)}^{VR}$				
			$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$		$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$	
1	I	0.596	0.039	0.203	0.281	0.108	0.454	0.040	0.223	0.258	0.112	0.406
	NI	0.619	0.110	0.205	0.295	0.135	0.457	0.092	0.235	0.273	0.136	0.435
2	I	0.468	0.377	0.385	0.418	0.379	0.436	0.391	0.398	0.415	0.392	0.416
	NI	0.474	0.375	0.378	0.414	0.374	0.438	0.392	0.396	0.413	0.391	0.423
3	I	0.439	0.400	0.403	0.420	0.401	0.432	0.414	0.417	0.425	0.415	0.415
	NI	0.443	0.400	0.401	0.418	0.399	0.435	0.416	0.416	0.425	0.415	0.420
$\infty$	I	0.417	0.418	0.418	0.418	0.418	0.431	0.431	0.431	0.432	0.431	0.418
	NI	0.418	0.418	0.418	0.419	0.418	0.432	0.432	0.432	0.433	0.432	0.418

Table 4.4 reports the simulation result on the average absolute relative bias ( $\overline{ARB}$ ) of MSE estimators under the Asparouhov size measures (4.3) and (4.4). It shows that  $\overline{ARB}$  of the MSE estimator of the EBLUP, based on the model without the augmenting variable  $g_{ij}$ , is very large when the sampling is very informative ( $\alpha = 1$ ): 52.8% for I and 59.1% for NI.  $\overline{ARB}$  gradually decreases to around 5% under non-informative sampling ( $\alpha = \infty$ ). The use of  $\log p_{j|i}$  as an augmenting variable leads to large reduction in  $\overline{ARB}$  ( $< 9\%$ ) and the choices  $p_{j|i}$  and  $n_i w_{j|i}$  also perform well in terms of  $\overline{ARB}$  except for the case of NI and  $\alpha = 1$  which leads to 18.5% and 12.9% respectively. Again,  $w_{j|i}$  is not a good choice because it leads to  $\overline{ARB}$  as large as 40% when  $\alpha = 1$ . The MSE estimator associated with the pseudo-EBLUP (without  $g_{ij}$ ) also performs well, except for NI and  $\alpha = 1$ , leading to  $\overline{ARB}$  of 19.5%. Use of  $\log p_{j|i}$  as auxiliary variable leads to  $\overline{ARB}$  less than 8% for the MSE estimator associated with the pseudo-EBLUP. We have not included the PS bootstrap MSE estimator in our study.

Overall, our simulation study indicates that the use of augmented models under informative sampling leads to EBLUPs that perform well in terms of  $\overline{AB}$  and  $\overline{RMSE}$  of the estimators, and  $\overline{ARB}$  of MSE estimators, provided that the augmenting variable is chosen properly. The bias-adjusted estimators of PS also perform well, even though they led to larger  $\overline{RMSE}$  under the PS size measures (4.2). Pseudo-EBLUP estimators (without the augmenting variable) also perform well and further improvement may be achieved under augmented models.

**Table 4.4**  
Average relative bias (%) of MSE estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

$\alpha$	Size measure	EBLUP					pseudo-EBLUP				
		$\hat{Y}_i^H$	$\hat{Y}_{i(\alpha)}^H$				$\hat{\mu}_i^{VR}$	$\hat{\mu}_{i(\alpha)}^{VR}$			
			$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$		$p_{j i}$	$n_i w_{j i}$	$w_{j i}$	$\log p_{j i}$
1	I	52.8	6.5	4.8	39.8	3.3	11.7	6.6	7.8	19.2	6.2
	NI	59.1	18.5	12.9	39.4	7.8	19.5	26.0	10.2	16.6	6.0
2	I	19.4	6.0	5.5	10.7	5.9	3.9	6.3	6.0	7.3	6.4
	NI	22.6	8.8	8.0	11.3	8.6	4.2	6.7	6.0	7.4	6.7
3	I	7.1	5.5	5.5	5.3	5.5	4.4	6.0	6.3	7.2	6.3
	NI	8.9	7.3	7.0	5.9	7.2	4.0	7.1	7.0	7.3	7.2
$\infty$	I	5.1	5.1	5.0	5.0	5.1	5.1	5.2	5.3	5.3	5.2
	NI	5.0	4.9	4.9	4.9	4.9	4.9	5.0	5.1	5.1	5.0

## 5 Concluding remarks

In this paper, we studied model-based small area estimation for different levels of design informativeness under a nested error linear regression model for the population units. Estimators considered were the EBLUP, the pseudo-EBLUP (You and Rao 2002) and an estimator given by Pfeffermann and Sverchkov (2007). The EBLUP and the pseudo-EBLUP were computed under two scenarios: (i) Ignore informative sampling and assume that the population model holds for the sample; (ii) Take account of informative sampling by using a suitable function of the unit selection probability  $p_{ji}$  as an additional auxiliary variable in the sample model.

Results from a simulation study showed that design informativeness can have a big impact on the bias and MSE of the EBLUP that ignores informative sampling (scenario (i)). Results under scenario (ii) showed that the EBLUP, based on the augmented model, performs extremely well in terms of bias and MSE, provided that the augmenting variable is chosen properly. The bias-adjusted estimator of Pfeffermann and Sverchkov (2007) also performed well under informative sampling in terms of bias but its MSE is significantly larger than the corresponding MSE of the EBLUP and the pseudo-EBLUP based on the augmented model. Pseudo-EBLUP under scenario (i) performed significantly better than the corresponding EBLUP. It can be significantly improved by using the augmented model, similar to the case of EBLUP.

An advantage of the augmented model approach is that no new theory is required for estimation and MSE estimation. However, the area mean  $\bar{G}_i$  of the augmenting variable  $g_{ij}$  is required, unlike in the approach of Pfeffermann and Sverchkov (2007). For some choices of  $g_{ij}$ ,  $\bar{G}_i$  is readily known; for example  $g_{ij} = p_{ji}$  gives  $\bar{G}_i = 1/N_i$  and  $g_{ij} = n_i w_{ji}$  gives  $\bar{G}_i = n_i \bar{W}_i$  and  $\bar{W}_i$  is often known for some surveys. We have also given a method of choosing the augmenting variable  $g_{ij}$ .

In this paper, we focused on the special case where all the areas are sampled. Extension of the augmented model approach to handle non-sampled areas requires the knowledge of the area means  $\bar{G}_i$ , as well as the area selection probabilities,  $p_i$ , for the non-sampled areas. This extension is currently under study.

## Acknowledgements

We are thankful to the Associate Editor, the referees and M. Sverchkov for many constructive comments and suggestions.

## References

- Asparouhov, T. (2006). General multi-level modelling with sampling weights. *Communication in Statistics, Theory and Methods*, 439-460.
- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- Pfeffermann, D., and Sverchkov, M. (2003). Small area estimation under informative sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3284-3295.
- 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. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 1, 67-72.
- Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- Rao, J.N.K. (2003). *Small Area Estimation*, New York: John Wiley & Sons, Inc.
- Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- Skinner, C.J. (1994). Sampling models and weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 133-142.
- Stefan, M. (2005). Contributions à l'estimation pour petits domaines. Ph.D. thesis, Université Libre de Bruxelles.
- Torabi, M., and Rao, J.N.K. (2010). Mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38, 4, 595-608.
- 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, 3, 431-439.