# SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING

*Danny Pfeffermann*, **Hebrew University and University of Southampton**
**and**
*Michail Sverchkov*, **Bureau of Labor Statistics and User Technology Associates, Inc.**

D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem, Israel 91905

**KEY WORDS: Prediction bias, Sample distribution, Sample-complement distribution, Sampling weights**

***** Opinions expressed in this paper are of the authors and do not constitute a policy of the Bureau of Labor Statistics.

## 1. Introduction

It is now generally accepted that small area estimation should be based on statistical models that permit borrowing information across areas or over time. See the recent book by Rao (2003) for a thorough discussion and comprehensive account of available methods. All the models and estimators considered so far assume either that all the population areas are represented in the sample or that the sampled areas are selected with equal probabilities. Only few studies consider the case where the sampling of units within the selected areas is with unequal selection probabilities, see e.g., Kott (1990), Arora and Lahiri (1997) and Prasad and Rao (1999). In this article we consider situations where the selection of the sampled areas is with probabilities that are related to the true (unknown) area means, and the sampling of units within the selected areas is with probabilities that are related to the study variable values, even when conditioning on the model covariates.

The problem with this kind of sampling designs is that the model holding for the sample data can differ from the model holding for the population values, giving rise to what is known in the sampling literature as '*informative sampling*'. As illustrated in this article, failure to account for the effects of an informative sampling scheme may result in severe bias of the small area predictors.

We use relationships between the *population distribution*, the *sample distribution* and the *sample-complement distribution* of a study variable developed in Pfeffermann and Sverchkov (1999) and Sverchkov and Pfeffermann (2000) (see next section), in order to derive approximately unbiased semi-parametric predictors of the area means under informative sampling schemes for both sampled and nonsampled areas. A fully parametric approach that consists of modeling the area means as functions of the area sample selection probabilities is also considered. Empirical results illustrating the biases that could be encountered when ignoring the sampling process and the performance of the proposed predictors are shown. We conclude with a brief summary that contains an outline for future research.

## 2. The sample and sample-complement distributions

Consider a finite population U consisting of *N* units belonging to *M* areas, with $N_i$ units in area $i$, $\sum_{i=1}^{M} N_i = N$. Let *y* define the study variable with value $y_{ij}$ for unit $j$ in area $i$ and denote by $x_{ij}$ the values of auxiliary (covariate) variables associated with that unit. In what follows we consider the population *y*-values as random outcomes of the following two level model:

*First level-* values (*random effects*) $\{u_1 ... u_M\}$ are generated independently from some distribution with probability density function (*pdf*) $f_p(u_i)$ for which $E_p(u_i) = 0$ ; $E_p(u_i^2) = \sigma_u^2$, where $E_p$ defines the expectation operator

*Second level-* values $\{y_{i1} ... y_{iN_i}\}$ are generated from some conditional distribution with *pdf* $f_p(y_{ij} \mid x_{ij}, u_i)$, for $i = 1 ... M$.

We assume a two-stage sampling scheme by which in the first stage *m* areas are selected with probabilities $\pi_i = \Pr(i \in s)$, and in the second stage $n_i$ units are sampled from area $i$ selected in the first stage with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$. Note that the sample inclusion probabilities at both stages may depend in general on all the population or area values of *y, x* and possibly also design variables *z*, used for the sample selection but not included in the working model. Also, the population areas are not necessarily the same as the primary sampling units (PSUs). Denote by $\mathrm{I}_i$ and $\mathrm{I}_{ij}$ the sample indicator variables at the two stages ($\mathrm{I}_i = 1$ iff $i \in s$ and similarly for $\mathrm{I}_{ij}$), and by $w_i = 1/\pi_i$ and $w_{j|i} = 1/\pi_{j|i}$ the corresponding first and second stage sampling weights.

Following Pfeffermann *et. al* (1998), we define the conditional *sample pdf* of $u_i$, i.e., the first level *pdf* of $u_i$ for area $i \in s$ as,

$$f_s(u_i) \overset{def}{=} f(u_i \mid \mathrm{I}_i = 1) \overset{Bayes}{=} \frac{\Pr(\mathrm{I}_i = 1 \mid u_i) f_p(u_i)}{\Pr(\mathrm{I}_i = 1)} \tag{2.1}$$

Similarly, the conditional *sample-complement pdf*, i.e., the conditional *pdf* of $u_i$ for area $i \notin s$ is defined in Sverchkov and Pfeffermann (2000) as,

$$f_c(u_i) \overset{def}{=} f(u_i \mid \mathrm{I}_i = 0) \overset{Bayes}{=} \frac{\Pr(\mathrm{I}_i = 0 \mid u_i) f_p(u_i)}{\Pr(\mathrm{I}_i = 0)} \tag{2.2}$$

Notice that the *population*, *sample* and *sample-complement pdf*s of $u_i$ are the same iff $\Pr(\mathrm{I}_i = 1 \mid u_i) = \Pr(\mathrm{I}_i = 1)\ \forall i$, in which case the sampling of areas is *noninformative*.

The second level *sample pdf* and *sample-complement pdf* of $y_{ij}$ are defined similarly to (2.1) and (2.2) as,

$$f_s(y_{ij} \mid x_{ij}, u_i) \overset{def}{=} f(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_{ij} = 1) = \frac{\Pr(\mathrm{I}_{ij} = 1 \mid y_{ij}, \mathbf{x}_{ij}, u_i) f_p(y_{ij} \mid \mathbf{x}_{ij}, u_i)}{\Pr(\mathrm{I}_{ij} = 1 \mid \mathbf{x}_{ij}, u_i)} \tag{2.3}$$

$$f_c(y_{ij} \mid x_{ij}, u_i) \overset{def}{=} f(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_{ij} = 0) = \frac{\Pr(\mathrm{I}_{ij} = 0 \mid y_{ij}, \mathbf{x}_{ij}, u_i) f_p(y_{ij} \mid \mathbf{x}_{ij}, u_i)}{\Pr(\mathrm{I}_{ij} = 0 \mid \mathbf{x}_{ij}, u_i)} \tag{2.4}$$

Here again the *population*, *sample* and *sample-complement pdf*s of $y_{ij}$ are the same iff $\Pr(\mathrm{I}_{ij} = 1 \mid y_{ij}, \mathbf{x}_{ij}, u_i) = \Pr(\mathrm{I}_{ij} = 1 \mid \mathbf{x}_{ij}, u_i)\ \forall j$, in which case the sampling of units within the selected areas is *noninformative*. The model defined by (2.1) and (2.3) defines the two-level *sample model* analogue of the population model defined by $f_p(u_i)$ and $f_p(y_{ij} \mid x_{ij}, u_i)$; see also Pfeffermann *et. al* (2001).

The following relationships between the three distributions are established in Pfeffermann and Sverchkov (1999) and Sverchkov and Pfeffermann (2000) for general pairs of random variables $v_1, v_2$ measured on elements $i \in U$, where $E_p, E_s$ and $E_c$ define respectively expectations under the *population*, *sample* and *sample-complement* distributions and $(\pi_i, w_i)$ denote the sample inclusion probabilities and the sampling weights.

$$f_s(v_{1i} \mid v_{2i}) = f(v_{1i} \mid v_{2i}, i \in s) = \frac{E_p(\pi_i \mid v_{1i}, v_{2i}) f_p(v_{1i} \mid v_{2i})}{E_p(\pi_i \mid v_{2i})} \tag{2.5}$$

$$E_p(v_{1i} \mid v_{2i}) = \frac{E_s(w_i v_{1i} \mid v_{2i})}{E_s(w_i \mid v_{2i})} \tag{2.6a}$$

$$E_p(\pi_i \mid v_{2i}) = \frac{1}{E_s(w_i \mid v_{2i})} \tag{2.6 b}$$

$$f_c(v_{1i}\,|\,v_{2i}) = f(v_{1i}\,|\,v_{2i}, i \notin s)$$
$$= \frac{E_p[(1-\pi_i)\,|\,v_{1i}, v_{2i}]\, f_p(v_{1i}\,|\,v_{2i})}{E_p[(1-\pi_i)\,|\,v_{2i}]} \qquad (2.7)$$
$$= \frac{E_s[(w_i-1)\,|\,v_{1i}, v_{2i}]\, f_s(v_{1i}\,|\,v_{2i})}{E_s[(w_i-1)\,|\,v_{2i}]}$$

$$E_c(v_{1i}\,|\,v_{2i}) = \frac{E_p[(1-\pi_i)v_{1i}\,|\,v_{2i}]}{E_p[(1-\pi_i)\,|\,v_{2i}]}$$
$$= \frac{E_s[(w_i-1)v_{1i}\,|\,v_{2i}]}{E_s[(w_i-1)\,|\,v_{2i}]} \qquad (2.8)$$

Defining $v_{1i} = u_i$, $v_{2i}$=*constant* yields the relationships holding for the random area effects $u_i$. Defining $v_{1ij} = y_{ij}$; $v_{2ij} = (x_{ij}, u_i)$ and substituting $\pi_{j|i}$ and $w_{j|i}$ for $\pi_i$ and $w_i$ respectively, yields the relationships holding for the observations $y_{ij}$.

## 3. Optimal Small Area Predictors

The target population parameters are the small area means $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ for $i = 1...M$, (the means in sampled and nonsampled areas). Let $D_s = \{(y_{ij}, w_{j|i}, w_i),\ (i,j) \in s;\ (\mathrm{I}_k, \mathrm{I}_{kl}, x_{kl}),\ (k,l) \in U\}$ define the known data. The MSE of a predictor $\hat{\bar{Y}}_i$ with respect to the *population pdf* given $D_s$ is,

$$MSE(\hat{\bar{Y}}_i \,|\, D_s) = E_p[(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \,|\, D_s]$$
$$= [\hat{\bar{Y}}_i - E_p(\bar{Y}_i \,|\, D_s)]^2 + V_p(\bar{Y}_i \,|\, D_s) \qquad (3.1)$$

The variance $V_p(\bar{Y}_i \,|\, D_s)$ does not depend on the form of the predictor and hence the MSE is minimized when $\hat{\bar{Y}}_i = E_p(\bar{Y}_i \,|\, D_s)$.

In what follows we distinguish between *sampled areas* ($\mathrm{I}_i = 1$) and *nonsampled areas* ($\mathrm{I}_i = 0$). Denote by $s_i$ the sample of units in sampled area $i$. Then, for the sampled areas,

$$E_p(\bar{Y}_i \,|\, D_s, \mathrm{I}_i = 1) = \frac{1}{N_i}[\sum_{j \in s_i} E_p(y_{ij} \,|\, D_s)$$
$$+ \sum_{l \notin s_i} E_p(y_{il} \,|\, D_s, \mathrm{I}_i = 1, \mathrm{I}_{il} = 0)] \qquad (3.2)$$
$$= \frac{1}{N_i}[\sum_{j \in s_i} y_{ij} + \sum_{l \notin s_i} E_c(y_{il} \,|\, D_s, \mathrm{I}_i = 1)]$$

For areas $i$ not in the sample,

$$E_p(\bar{Y}_i \,|\, D_s, \mathrm{I}_i = 0) = \frac{1}{N_i}\sum_{k=1}^{N_i} E_p(y_{ik} \,|\, D_s, \mathrm{I}_i = 0)$$
$$= \frac{1}{N_i}\sum_{k=1}^{N_i} E_c(y_{ik} \,|\, D_s, \mathrm{I}_i = 0) \qquad (3.3)$$

The predictors in (3.2) and (3.3) can be written in a single equation as,

$$E_p(\bar{Y}_i \,|\, D_s) = \frac{1}{N_i}\{\sum_{k=1}^{N_i} y_{ik}\mathrm{I}_{ik} +$$
$$\sum_{k=1}^{N_i} E_c[y_{ik}(1-\mathrm{I}_{ik}) \,|\, D_s, \mathrm{I}_i = 1]\}\mathrm{I}_i \qquad (3.4)$$
$$+ \frac{1}{N_i}\sum_{k=1}^{N_i} E_c[y_{ik} \,|\, D_s, \mathrm{I}_i = 0](1-\mathrm{I}_i)$$

## 4. Bias of Small Area Predictors when ignoring the Sampling Scheme

Consider for convenience the case of a sampled area. Ignoring the sampling scheme implies an implicit assumption that the *sample-complement* model is the same as the *sample* model, such that $\hat{\bar{Y}}_{i,IGN} = \frac{1}{N_i}[\sum_{j \in s_i} y_{ij} + \sum_{l \notin s_i} E_s(y_{il} \,|\, D_s, \mathrm{I}_i = 1)]$ where $E_s(y_{il} \,|\, D_s, \mathrm{I}_i = 1)$ defines the expectation of *y* with respect to the sample distribution. Hence,

$$E_p[(\hat{\bar{Y}}_{i,IGN} - \bar{Y}_i) \,|\, D_s, \mathrm{I}_i = 1]$$
$$= \frac{1}{N_i}\sum_{l \notin s_i} E_s(y_{il} \,|\, D_s, \mathrm{I}_i = 1)$$
$$- \frac{1}{N_i}\sum_{l \notin s_i} E_c(y_{il} \,|\, D_s, \mathrm{I}_i = 1)$$
$$= -\frac{1}{N_i}\sum_{l \notin s_i} \frac{Cov_s(y_{il}, w_{l|i} \,|\, D_s, \mathrm{I}_i = 1)}{E_s[(w_{l|i}-1) \,|\, D_s, \mathrm{I}_i = 1]} \qquad (4.1)$$

with the second equality following from (2.8). Thus, unless the study values $y_{il}$ and the sampling weights $w_{l|i}$ within the selected areas are uncorrelated, ignoring the sampling scheme results in biased predictors (see also the empirical results). A similar expression of the bias can be obtained for nonsampled areas.

Example

Let the *population model* be the 'unit level random effects model'

$$y_{ij} = \mu + u_i + e_{ij} \ ;$$

$$u_i \sim N(0, \sigma_u^2) \ , \ e_{ij} \sim N(0, \sigma_e^2) \qquad (4.2)$$

with the random effects and residual terms being mutually independent. Consider the common sampling scheme by which *m* areas are sampled with probabilities $\pi_i = c \times N_i$ for some constant *c*, and $\pi_{j|i} = n_0 / N_i$ (fixed sample size $n_0$ within the selected areas), such that $\pi_{ij} = \Pr[(i,j) \in s] = \pi_i \pi_{j|i} = const$. (For fixed *m*, $c = m/N$.) Note that sampling within the selected areas is *noninformative* in this case but if the area sizes $N_i$ are correlated with the random effects $u_i$, the selection of the areas is *informative* (say, the areas are *school districts*, the study variable measures children's attainments, the large districts are rich areas with high school attainments).

Suppose that the area sizes can be modeled as $\log(N_i) \sim N(Au_i + B, \sigma_M^2)$ with *A*>0, implying that,

$$E_p(\pi_i \mid u_i) \prec \exp(Au_i + B + \frac{\sigma_M^2}{2}).$$ It follows that (see Pfeffermann *et al*. 1998 example 4.3),

$$f_s(u_i) = \frac{E_p(\pi_i | u_i) f_p(u_i)}{E_p(\pi_i)} = N(A\sigma_u^2, \sigma_u^2) \qquad (4.3)$$

so that $E_s(u_i) = A\sigma_u^2 \neq E_p(u_i) = 0$. The fact that the random effects in the sample have in this case a positive expectation is easily explained by the fact that the sampling scheme considered tends to select the areas with large positive random effects. Note, however, that by defining $\mu^* = \mu + A\sigma_u^2$ and $u_i^* = u_i - A\sigma_u^2$, the model holding for the sample data in sampled areas is $y_{ij} = \mu^* + u_i^* + e_{ij}$, $u_i^* \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, which is the same as the population model. Thus, the optimal predictors under the *population model* for the area means $\theta_i = \mu + u_i$ in sampled areas ($\mathrm{I}_i = 1$), are still optimal under the *sample model*. (Recall that for the present example the sampling scheme within the selected areas is noninformative.)

Next consider *nonsampled areas*. By (2.7),

$$f_c(u_i) = \frac{E_p[(1-\pi_i) \mid u_i] f_p(u_i)}{E_p(1-\pi_i)} = \frac{f_p(u_i)}{E_p(1-\pi_i)} - \frac{E_p(\pi_i \mid u_i) f_p(u_i)}{E_p(1-\pi_i)} \qquad (4.4)$$

Let

$$E_p(m) = E_p[\sum_{l=1}^{M} \mathrm{I}_l] = E_p[E_p(\sum_{l=1}^{M} \mathrm{I}_l \mid \{N_i\})] = E_p[\sum_{l=1}^{M} \pi_i] = ME_p(\pi_i)$$ define the expected number of sampled areas, such that $E_p(\pi_i) = E_p(m)/M$. If the number of sampled areas is fixed, $E_p(m) = m$. By (4.4) and (2.5),

$$f_c(u_i) = \frac{Mf_p(u_i) - E_p(m) f_s(u_i)}{M - E_p(m)}$$ and hence,

$$E_c(u_i) = -\frac{E_p(m) E_s(u_i)}{M - E_p(m)} = -\frac{E_p(m) A\sigma_u^2}{M - E_p(m)} \qquad (4.5)$$

Here again the negative expectation of the random effects pertaining to *nonsampled areas* is easily explained by the tendency of the sampling scheme to sample the areas with large positive random effects. It follows form (4.5) that ignoring the sampling scheme underlying the selection of the areas and predicting, for example, the sample means in nonsampled areas by the average of the predictors in the sampled areas yields in this example biased predictors with bias,

$$B(i \notin s) = A\sigma_u^2 - [-A\sigma_u^2 \frac{E_p(m)}{M - E_p(m)}] = A\sigma_u^2 \frac{M}{M - E_p(m)} \qquad (4.6)$$

## 5. Inference

The first step of the proposed approach is to fit a model to the sample data, which of course is a necessary step in any small area estimation problem. Notice that although we consider informative sampling, the sample model can be identified from the sample data using standard techniques, see, e.g., Rao (2003) for details.

In what follows we suppose therefore that an adequate model has been fitted to the sample data, and in order to illustrate the estimation procedures we assume that this model is the 'nested error regression model',

$$y_{ij} = x_{ij}'\beta + u_i + e_{ij} \;;\; u_i \mid \mathrm{I}_i = 1 \sim N(0, \sigma_u^2),$$
$$e_{ij} \mid \mathrm{I}_{ij} = 1 \sim N(0, \sigma_e^2) \qquad (5.1)$$

The model defined by (5.1) is in common use for small area estimation problems, see e.g., Battesse *et al.* (1988). Suppose further that the sampled areas had been included in the sample with inclusion probabilities $\pi_i$, $i = 1 \ldots m$, and that $n_i$ units were sampled from area $i$ in the sample with probabilities $\pi_{j|i}$. Finally, we assume that,

$$E_s(w_{j|i} \mid x_{ij}, y_{ij}, u_i) = E_s(w_{j|i} \mid x_{ij}, y_{ij})$$
$$= c_i \exp(a x_{ij} + b y_{ij}) \qquad (5.2)$$

where $c_i > 0$, $a$ and $b$ are fixed parameters.

<u>*Comment*</u>: As with the sample model (5.1), the expectation in (5.2) refers to the *sample distribution* within the areas. The relationship in the sample between the sampling weights and the observed data can be identified and estimated therefore from the sample data, see Pfeffermann and Sverchkov (1999, 2003) for discussion and examples. On the other hand, the relationship between the sampling weights $w_i$ and the area means is more difficult to detect since the area means are not observable, and in the rest of this section we do not model this relationship. See Section 6 and also Pfeffermann *et al.* (2001) for examples of modeling the area selection probabilities. Kim (2003) assumes the model (5.1) for the population values and a similar model to (5.2) for the sampling probabilities within the areas but it is assumed that all the areas in the population are represented in the sample.

As established in Section 3, the optimal predictor for a *sampled* area $i$ is,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 1) =$$

$$[\sum_{j \in s_i} y_{ij} + \sum_{l \notin s_i} E_c(y_{il} \mid D_s, \mathrm{I}_i = 1)] / N_i .$$

In order to compute the expectations $E_c(y_{il} \mid D_s, \mathrm{I}_i = 1)$ we proceed as follows: First, by (2.7), (5.1) and (5.2),

$$f_c(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)$$
$$= \frac{[E_s(w_{l|i} \mid x_{il}, y_{il}, u_i) - 1] f_s(y_{il} \mid x_{il}, u_i)}{E_s(w_{l|i} \mid x_{il}, u_i) - 1}$$
$$= \frac{\lambda_{il}}{\lambda_{il} - 1} \frac{1}{\sigma_e} \phi\left(\frac{y_{ij} - u_{il} - b\sigma_e^2}{\sigma_e}\right)$$
$$- \frac{1}{\lambda_{il} - 1} \frac{1}{\sigma_e} \phi\left(\frac{y_{il} - u_{il}}{\sigma_e}\right) \qquad (5.3)$$

where $u_{il} = x_{il}'\beta + u_i$,

$\lambda_{il} = c_i \exp[(b^2\sigma_e^2/2) + a x_{il} + b u_{il}]$

$= E_s(w_{l|i} \mid x_{il}, u_i)$ and $\phi$ is the standard normal *pdf*. Notice that if $b = 0$ (selection probabilities within the sampled areas only depend on the *x*-values and hence the sampling is noninformative), the *pdf* in (5.3) reduces to the conditional normal density defined by (5.1). Second, by (5.3),

$$E_c(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)$$
$$= u_{il} + \frac{\lambda_{il}}{\lambda_{il} - 1} b\sigma_e^2 \qquad (5.4)$$

Finally,

$$E_c(y_{il} \mid D_s, \mathrm{I}_i = 1)$$
$$= E_s[E_c(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1)] \qquad (5.5)$$
$$= E_s[E_c(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)]$$

where the exterior expectation is with respect to the distribution of $u_i \mid D_s, \mathrm{I}_i = 1$. Under the model (5.1), the latter distribution is known to be normal with mean $\hat{u}_i = \gamma_i[\bar{y}_i - \bar{x}_i'\beta]$ and variance $\sigma_i^2\gamma_i$, where $(\bar{y}_i, \bar{x}_i) = \sum_{j=1}^{n_i}(y_{ij}, x_{ij})/n_i$ are the sample means of (*y,x*) in sampled area $i$, $\sigma_i^2 = \sigma_e^2/n_i = Var(\bar{y}_i \mid u_i)$ and $\gamma_i = \sigma_u^2/[\sigma_u^2 + \sigma_i^2]$.

Thus, for the *sampled areas* $E_c(y_{il} \mid D_s, \mathrm{I}_i = 1)$ is obtained by computing the expectation of the right hand side of (5.4) with respect to the normal distribution of $u_i \mid D_s, \mathrm{I}_i = 1$. We find that,

$$E_c(y_{il} \mid D_s, \mathrm{I}_i = 1) = (x_{il}'\beta + \hat{u}_i) + b\sigma_e^2 E_s(\frac{1}{1-\lambda_{il}^{-1}} \mid D_s, \mathrm{I}_i = 1) \quad (5.6)$$

Notice that if $b$=0 (noninformative sampling within the areas) $E_c(y_{il} \mid D_s, \mathrm{I}_i = 1) = x_{il}'\beta + \hat{u}_i$, which is the standard result.

The expectation $E_s(\frac{1}{1-\lambda_{il}^{-1}} \mid D_s, \mathrm{I}_i)$ can be computed numerically. Alternatively, for the practical case where the sampling fractions within the selected areas are very small, $\lambda_{il} = E_s(w_{l\mid i} \mid x_{il}, u_i)$ is under mild conditions much larger than 1 and therefore we may approximate,

$$E_s(\frac{1}{1-\lambda_{il}^{-1}} \mid D_s, \mathrm{I}_i) \cong E_s[(1+\lambda_{il}^{-1}) \mid D_s, \mathrm{I}_i].$$

The latter expectation can be computed analytically yielding,

$$E_c(y_{il} \mid D_s, \mathrm{I}_i = 1) = \hat{u}_{il} + b\sigma_e^2[1 + \frac{1}{c_i}\exp(-\frac{b\sigma_e^2}{2} - ax_{il} - b\hat{u}_{il} + \frac{b^2\sigma_i^2\gamma_i}{2})] \quad (5.7)$$

where $\hat{u}_{il} = x_{il}'\beta + \hat{u}_i$.

It follows from (3.2) and (5.7) that for given parameters $\{\beta, c_i, a, b, \sigma_u^2, \sigma_e^2\}$, the optimal predictor of $\bar{Y}_i$ for *sampled* area $i$ is,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 1) = \frac{1}{N_i}\{(N_i - n_i)\hat{\theta}_i + n_i[\bar{y}_i + (\bar{X}_i - \bar{x}_i)'\beta] + (N_i - n_i)b\sigma_e^2 + \frac{b\sigma_e^2}{c_i}\exp(-\frac{b\sigma_e^2}{2} - b\hat{u}_i + \frac{b^2\sigma_i^2\gamma_i}{2})\sum_{l \notin s_i}\exp(-ax_{il} - bx_{il}'\beta)\} \quad (5.8)$$

where $\hat{\theta}_i = \hat{u}_i + \bar{X}_i\beta$ is the optimal predictor of the area mean $\theta_i = \bar{X}_i'\beta + u_i = E_s(\bar{Y}_i \mid u_i)$. The terms in (5.8) that are multiplied by $b$ correct for the difference between the sample-complement expectation and the sample expectation. Notice on the other hand that even under noninformative sampling ($b$=0), the predictor implied by (5.8) differs from the predictor $\hat{\theta}_i$ in common use. This is so because the target parameter is defined to be the finite area mean $\bar{Y}_i$ rather than $\theta_i$, see Rao (2003, Eq. 7.2.37).

For the *nonsampled* areas, the optimal predictor of the area means is defined in (3.3) to be, $E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 0) = \sum_{k=1}^{N_i} E_c(y_{ik} \mid D_s, \mathrm{I}_i = 0)/N_i$. By (2.8) and then (2.6),

$$E_c(y_{ik} \mid D_s, \mathrm{I}_i = 0) = E_c[E_p(y_{ik} \mid x_{ik}, u_i) \mid D_s] = E_s[\frac{(w_i - 1)E_p(y_{ik} \mid x_{ik}, u_i)}{E_s(w_i \mid D_s) - 1} \mid D_s] = \frac{E_s[(w_i - 1)\frac{E_s(w_{k\mid i}\, y_{ik} \mid x_{ik}, u_i)}{E_s(w_{k\mid i} \mid x_{ik}, u_i)} \mid D_s]}{E_s(w_i \mid D_s) - 1} \quad (5.9)$$

Computing the two expectations in the numerator of the last expression of (5.9) employing (5.1) and (5.2), yields after some algebra,

$$E_c(y_{ik} \mid D_s, \mathrm{I}_i = 0) = x_{ik}'\beta + b\sigma_e^2 + E_s[\frac{(w_i - 1)u_i}{E_s(w_i \mid D_s) - 1} \mid D_s] \quad (5.10)$$

Estimating $u_i = E_s[(y_{ij} - x_{ij}'\beta) \mid u_i]$ and $E_s(w_i \mid D_s)$ by the corresponding sample means $\hat{u}_{S,i} = \sum_{j \in s_i}(y_{ij} - x_{ij}'\beta)/n_i$ and $\hat{E}_s(w_i \mid D_s) = \sum_{i \in s} w_i / m$ (application of the method of moments), and substituting the estimates in (5.10), gives the following simple estimate for $E_c(y_{ik} \mid D_s, \mathrm{I}_i = 0)$,

$$\hat{E}_c(y_{ik} \mid D_s, \mathrm{I}_i = 0) = x_{ij}'\beta + b\sigma_e^2 + \frac{\sum_{i \in s}(w_i - 1)\hat{u}_{S,i}}{\sum_{i \in s}(w_i - 1)} \quad (5.11)$$

It follows from (3.3) and (5.10) that for given parameters $\{\beta, c_i, a, b, \sigma_u^2, \sigma_e^2\}$, the optimal predictor of $\bar{Y}_i$ for area $i$ not in the sample is,

$$\hat{E}_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 0) = \bar{X}_i'\beta + b\sigma_e^2 + \frac{\sum_{i\in s}(w_i - 1)\hat{u}_{S,i}}{\sum_{i\in s}(w_i - 1)} \quad (5.12)$$

## 6. Parametric estimation

The analysis of the preceding section makes no assumptions regarding the form of the relationship between the area selection probabilities and the area means. However, in situations where this relationship can be modeled adequately, more efficient predictors can be obtained by incorporating the area selection probabilities into the model, thus making the sampling scheme noninformative.

For example, suppose that the population model is the 'unit level random effect model'

$$y_{ij} = \mu + u_i + e_{ij}$$
$$u_i \sim N(0, \sigma_u^2)\ ;\ e_{ij} \sim N(0, \sigma_e^2)$$
$$i = 1...M\ ,\ j = 1...N_i \quad (6.1)$$

Suppose further that the area sizes can be modeled as, $N_i = Int[K_1 \times \exp(K_2 u_i)]$ where $K_1$ and $K_2$ are constants, and that the selection of the areas is with probabilities $\pi_i$ that are proportional to the sizes $N_i$. Then, to a close approximation, the distribution of $u_i \mid \log(\pi_i)$ is normal and the population model can be written as,

$$y_{ij} = \mu + u_i + e_{ij}$$
$$= \beta_0 + \beta_1 \log(\pi_i) + \eta_i + e_{ij}\ ; \quad (6.2)$$
$$\eta_i \sim N(0, \sigma_\eta^2)\ ,\ e_{ij} \sim N(0, \sigma_e^2)$$

For the case where the selection of the samples within the sampled areas is with equal probabilities, the population model defined by (6.2) holds also for the sample data (the sampling scheme is noninformative for this model), and the small area means can be predicted using standard procedures. Thus, by incorporating the area selection probabilities into the model, the sampling scheme becomes noninformative. Notice, however, that the use of this model requires knowledge of the area selection probabilities for all $i = 1...M$. See Rubin (1985) and Skinner (1994) for more general results and discussion of the implications of including the selection probabilities among the covariates of statistical models.

When the selection of the samples within the sampled areas is informative, one can either model the relationship between the selection probabilities and the study variable and apply the methodology described in Section 5, (see also Kim 2003, this article assumes that all the population areas are represented in the sample), or use the weighted direct estimates of the area means as the observed data. (In practice, the only available data to the analyst are often the direct estimates anyway.) See Kott (1990), Arora and Lahiri (1997) and Prasad and Rao (1999) for modeling of the direct area estimates. All these studies, however, assume implicitly noninformative sampling.

Suppose that the direct estimates are the Hajek's estimates,

$$\hat{\theta}_{H,i} = \sum_{j\in s_i} w_{j|i}\, y_{ij} \Big/ \sum_{j\in s_i} w_{j|i} \quad (6.3)$$

where $\theta_i = \mu + u_i = E_p(\bar{Y}_i \mid u_i)$ defines the $i$-th area mean. Then, under (6.2), combined with the common assumption that $\hat{\theta}_{H,i} = \theta_i + \zeta_i$ , $\zeta_i \sim N(0, \sigma_\zeta^2)$ where $\zeta_i$ defines the sampling error in area $i$, the model holding for the direct estimators is,

$$\hat{\theta}_{H,i} = \theta_i + \zeta_i\ ,\ \zeta_i \sim N(0, \sigma_\zeta^2)\ ;$$
$$\theta_i = \beta_0 + \beta_1 \log(\pi_i) + \eta_i$$
$$\eta_i \sim N(0, \sigma_\eta^2) \quad (6.4)$$

The model defined by (6.4) is the familiar Fay and Herriot (1979) model that is in common use for small area estimation. Notice again that this model can be fitted using standard procedures (ignoring the sampling process).

## 7. Monte-Carlo simulation study

In order to illustrate the biases that could occur when ignoring an informative sampling scheme (employing standard predictors), and to study the performance of the predictors proposed in this article that account for the sampling process, we designed a small simulation study. The study was carried out as follows:

1- Generate population random area effects $u_i \sim N(0,\sigma_u^2)$ and area sizes $N_i = Int[1000\times\exp[u_i/(5\sigma_u)]$ for $i=1...M$ ($\sigma_u^2=16$, $M$=150 population areas).

2- Generate $y$-values using the model (6.1) with $\mu=20$, $\sigma_e^2=100$.

3- Select areas with probabilities $\pi_i = mN_i/\sum_{j=1}^{150}N_i$ ($m$=90 sampled areas) using Systematic PPS sampling.

4- Sample $n_i$ units from selected area $i$ with probabilities $\pi_{j|i} = n_0 z_{ij}/\sum_{k=1}^{N_i} z_{ik}$ ($n_i = n_0 = 5$ sampled units in each selected area), where $z_{ij}=\exp(y_{ij}/50)$, again using Systematic PPS sampling. These selection probabilities satisfy the relationship (5.2) with $a=0$, $b=-(1/50)$ and $c_i = (N_i/n_0)E(z_{ij})$.

Repeat Steps 1-4 1000 times.

For each sample we computed the following 4 predictors of the area means:

A- $\hat{\theta}_i = \hat{\gamma}_i\bar{y}_i + (1-\hat{\gamma}_i)\bar{y}$ for sampled areas, $\hat{\theta}_i = \sum_{i\in s}[\hat{\gamma}_i\bar{y}_i + (1-\hat{\gamma}_i)\bar{y}]/m = \bar{y}$ for nonsampled areas; $\bar{y} = \sum_{i\in s}\bar{y}_i/m$, $\hat{\gamma}_i = \hat{\sigma}_u^2/[\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_0]$, $(\hat{\sigma}_u^2,\hat{\sigma}_e^2)$ computed by method of moments; see Prasad and Rao (1990).

The predictors $\{\hat{\theta}_i\}$ are the ordinary predictors under noninformative sampling (ignoring the sample selection), when the area sizes are sufficiently large such that $\sum_{k=1}^{N_i} e_{ij}/N_i \cong 0$

B- $\hat{\theta}_{H,i}$ (Eq. 6.3) for sampled areas ('direct estimator'), $\tilde{\theta}_{H,i} = \sum_{i\in s} w_i\hat{\theta}_{H,i}/\sum_{i\in s} w_i$ for nonsampled areas.

C- Proposed 'semi-parametric' predictors assuming the relationship (5.2) for the weights $w_{j|i}$; The predictors are defined by (5.8) for the sampled areas and by (5.12) for the nonsampled areas with $a=0$, $x_{ij}=1$, $\beta=\mu$. Note that under the model (6.1) and the sampling scheme used to select the areas, the sample random effects also have a normal distribution but with a different expectation, thus justifying the use of the predictors (5.8) and (5.12). See Pfeffermann *et. al* (1998). The unknown model parameters have been replaced by sample estimates: $(\hat{\sigma}_u^2,\hat{\sigma}_e^2)$ computed by method of moments, $\hat{\mu}=\bar{y}$; the parameters $b$ and $c_i$ indexing the relationship between the weights $w_{j|i}$ and the study variable (Eq. 5.2) were estimated by fitting the model $w_{j|i} = c_i\exp(by_{ij})+\varepsilon_{ij}$, using the REG and NLIN procedures of SAS.

D- Fay-Herriot predictors obtained by fitting the model (6.4) (parametric approach); $\sigma_\eta^2$ estimated by the Fay-Herriot (1979) method, $\sigma_\zeta^2$ taken as known (computed as, $\hat{\sigma}_\zeta^2 = \sum_{i\in s}(\hat{\theta}_{H,i}-\bar{Y}_i)^2/m$). The assumption that the sampling error variance is known is common. Estimating the variance from the sample data yields very similar results. $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1)$ computed by generalized least squares with $\hat{\theta}_{H,i}$ as the dependent variable values and $(\sigma_\eta^2,\sigma_\zeta^2)$ replaced by $(\hat{\sigma}_\eta^2,\hat{\sigma}_\zeta^2)$.

The results of the simulation study are displayed in Figures 1-4. Figures 1 and 2 show for each area the empirical prediction bias and root mean square error (RMSE) of the four predictors over all the simulations for which that area has been sampled, Figures 3 and 4 show for each area the bias and RMSE of the four predictors over all the simulations for which that area has not been sampled. Denote by $\bar{Y}_{tr}$ the true area mean in simulation $r$, $r$=1…1000 and let $\hat{\bar{Y}}_{tr}$ represent any of the predictors. Define $D_{tr}=1$ if area $t$ was sampled in simulation $r$ and $D_{tr}=0$ otherwise. For a given area $t$, the prediction bias and RMSE when this area has been sampled is computed as,

$$Bias_t = \frac{\sum_{r=1}^{1000} D_{tr}(\hat{\bar{Y}}_{tr} - \bar{Y}_{tr})}{\sum_{r=1}^{1000} D_{tr}};$$

$$RMSE_t = \sqrt{\frac{\sum_{r=1}^{1000} D_{tr}(\hat{\bar{Y}}_{tr} - \bar{Y}_{tr})^2}{\sum_{r=1}^{1000} D_{tr}}} \quad (7.1)$$

The prediction bias and RMSE when area *t* has not been sampled are obtained by replacing $D_{tr}$ by $(1 - D_{tr})$ in (7.1). The four Figures show also at the bottom for each predictor the average of the corresponding measure over all the 150 areas.

The conclusions from this simulation study are clear-cut and can be summarized as follows:

1- Ignoring an informative sampling scheme can result in severe prediction bias for both sampled and nonsampled areas.

2- The direct (design based) estimators are approximately unbiased in sampled areas but are biased for the area means of nonsampled areas. This result is explained by the fact that the estimator $\tilde{\theta}_{H,i}$ estimates the average of the area means in the population, which is different from the average of the nonsampled area means. For the present population model under which the true area means $\theta_i$ are exchangeable, the predictor, $\hat{\theta}_{H,i} = \sum_{i \in s}(w_i - 1)\hat{\theta}_{H,i} / \sum_{i \in s}(w_i - 1)$ is approximately design unbiased for the area means in nonsampled areas, but this property does not necessarily hold for other models. In fact, no approximately design unbiased predictors for the area means in nonsampled areas exist in general.

3- The use of the semi-parametric approach yields unbiased predictors for both the sampled and nonsampled areas but with large RMSEs for nonsampled areas compared to the use of the fully parametric approach. Notice in this respect that under the present model and sampling scheme used for the selection of the areas, $E_c(\theta_i) = \mu - [m/(M-m)]\sigma_u / 5$ and $Var_c(\theta_i) = \sigma_u^2 - [mM/(M-m)^2](\sigma_u / 5)^2$. (The parameter $\sigma_u / 5$ indexes the distribution of the area sizes. See step one of the design of the simulation study.) For $\mu = 20,\ M = 150,\ m = 90,\ \sigma_u^2 = 16$ as in the present study, $E_c(\theta_i) = 18.8$ and $Var_c(\theta_i) = 13.6$. Thus, by predicting the area means of nonsampled areas by their true expected value of 18.8, the RMSE of the prediction error is $\sqrt{13.6} = 3.69$, which is only slightly lower than the average RMSE of 3.79 achieved by use of the semi-parametric predictors (see bottom of Figure 4). However, the use of the expected values under the model or the semi-parametric approach do not account for the relationship between the area means and the sample selection probabilities. As illustrated in the present study, including this relationship as part of the model can reduce the RMSE quite substantially, particularly for nonsampled areas where no direct estimates exist.

## 8. Summary

This article proposes two approaches for small area estimation under informative sampling. The semi-parametric approach makes no assumptions regarding the relationship between the area selection probabilities and the true area means. The proposed predictors under this approach are approximately unbiased for both the sampled and nonsampled areas but the prediction RMSEs can be large particularly for nonsampled areas.

The parametric approach models the relationship between the area selection probabilities and the area means and incorporates this relationship into the model for the study variable. As illustrated by the simulation study, the use of this approach can reduce the RMSEs quite significantly but a major issue that needs to be investigated is the robustness of the parametric predictors to misspecification of the model relating the area selection probabilities to the area means.

Two other outstanding issues are the development of appropriate MSE estimators for the nonparametric approach and the extension of this approach to other plausible sample models.

## References


Arora, V. and Lahiri, P. (1997), "On the superiority of the Bayesian method over the BLUP in small area estimation problems," *Statistica Sinica* **7**, 1053-1063.

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

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

Kim, D. H. (2002), "Bayesian and empirical Bayesian analysis under informative sampling," *Sankhya B*, **64**, 267-288.

Kott, P.S. (1990), "Robust small domain estimation using random effects modeling," *Survey Methodology* **15**, 3-12.

Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998), "Parametric distributions of complex survey data under informative probability sampling," *Statistica Sinica* **8**, 1087-1114.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P. L. (2001), "Multilevel modeling under informative sampling. *2001 Proceedings of the International Statistical Institute*, International Association of Survey Statisticians (IASS) Invited Paper Sessions, 505-532.

Pfeffermann, D. and Sverchkov, M. (1999), "Parametric and semi-parametric estimation of regression models fitted to survey data," Sankhya **61**, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003), "Fitting generalized linear models under informative probability sampling," In: *Analysis of Survey Data*, eds. C. J. Skinner and R. L. Chambers, New York: Wiley, 175-195.

Prasad, N. G. N and Rao, J. N. K. (1999), "On robust small area estimation using a simple random effects model," *Survey Methodology* **25**, 67-72.

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

Rubin D. B. (1985), "The use of propensity scores in applied Bayesian inference," In: *Bayesian Statistics 2*, eds. J. M. Bernando, M. H. Degroot, D. V. Lindley and A. F. M. Smith, Amsterdam: Elsevier Science, 463-472.

Skinner, C. J. (1994), "Sample models and weights," *1994 Proceedings of the American Statistical Association*, Survey Research Methods Section, 133-142.

Sverchkov, M. and Pfeffermann, D. (2000), "Prediction of finite population totals under informative sampling utilizing the sample distribution," *2000 Proceedings of the American Statistical Association*, Survey Research Methods Section, 41-46.

Figure 1. Prediction Bias in Sampled Areas

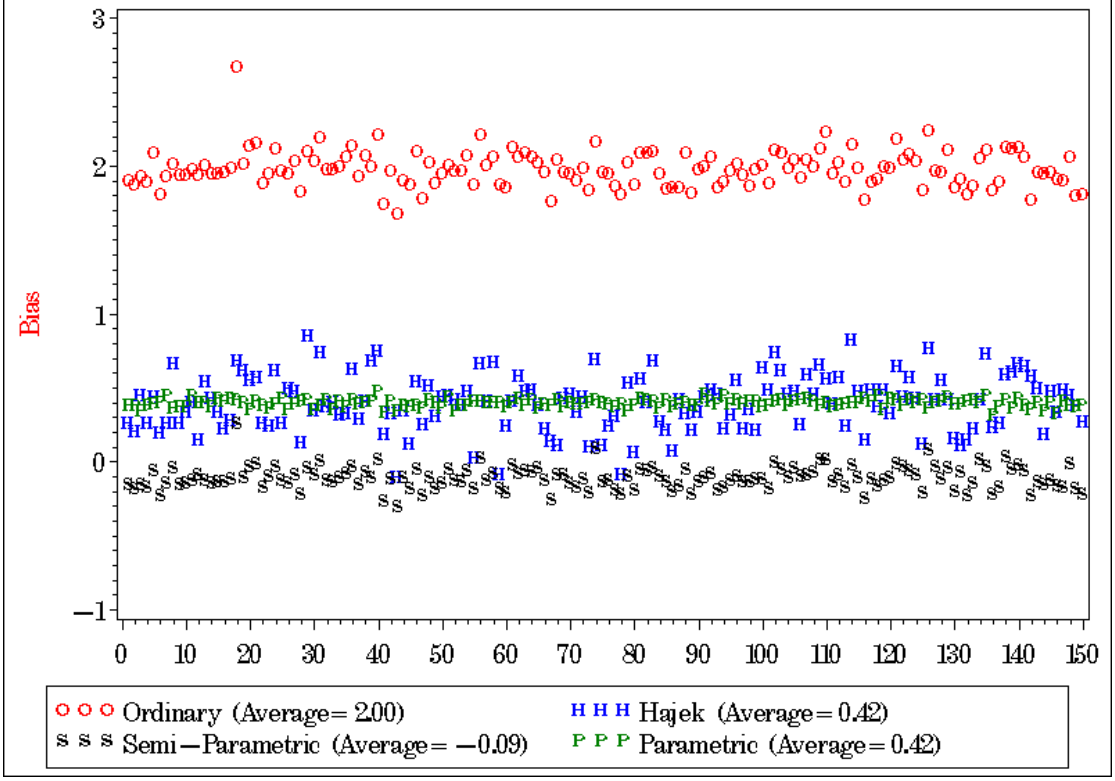
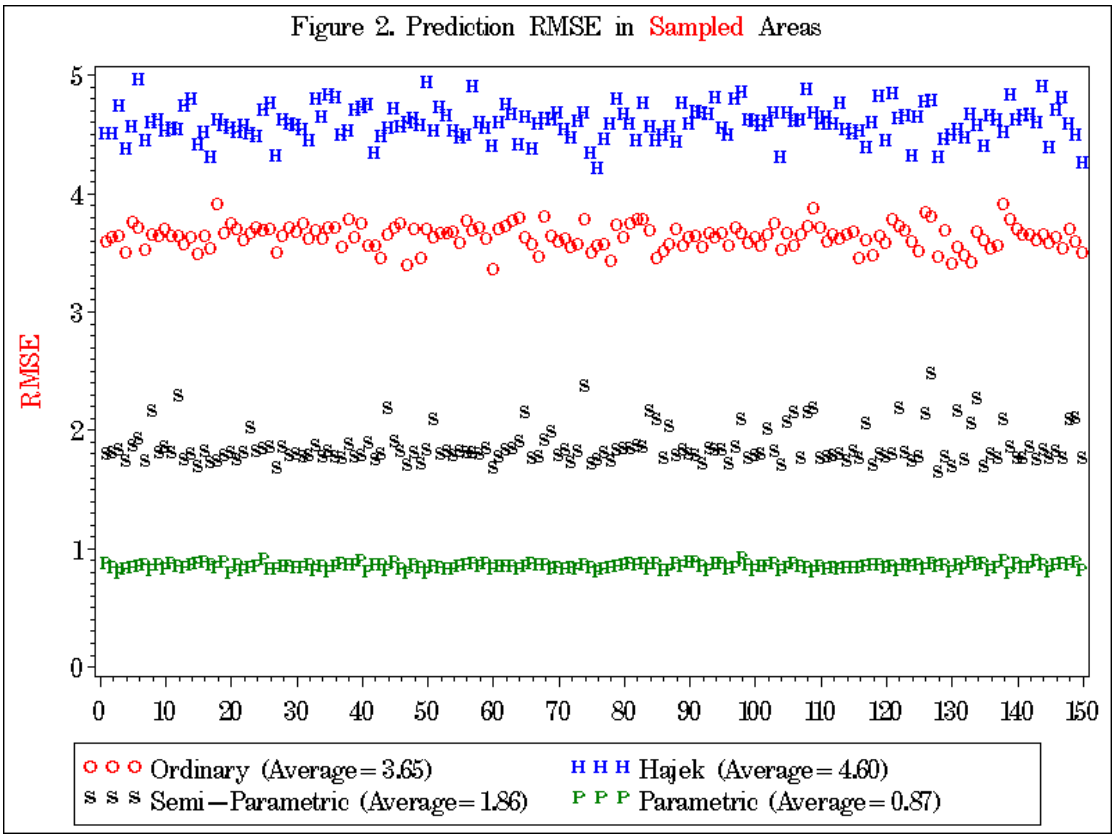



Figure 2. Prediction RMSE in Sampled Areas



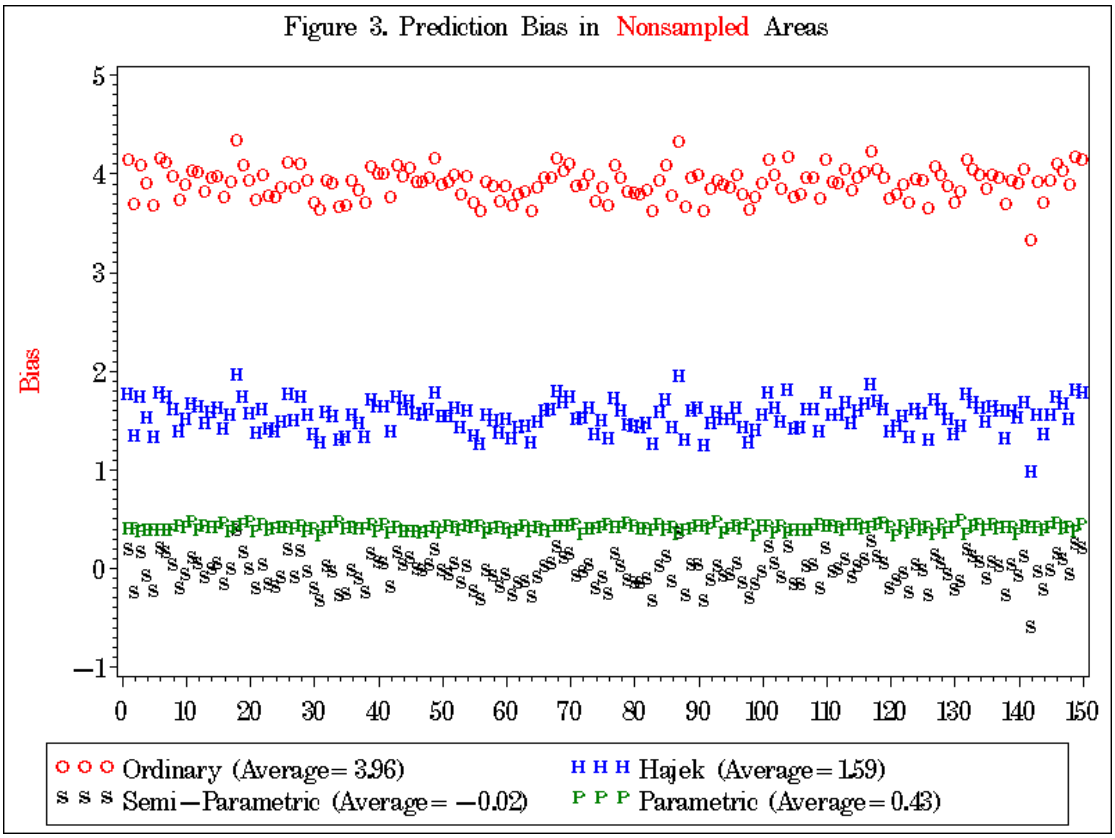


Figure 3. Prediction Bias in Nonsampled Areas
Bias
5
4
3
2
1
0
−1
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Ordinary (Average= 3.96)
Hajek (Average= 1.59)
Semi−Parametric (Average= −0.02)
Parametric (Average= 0.43)

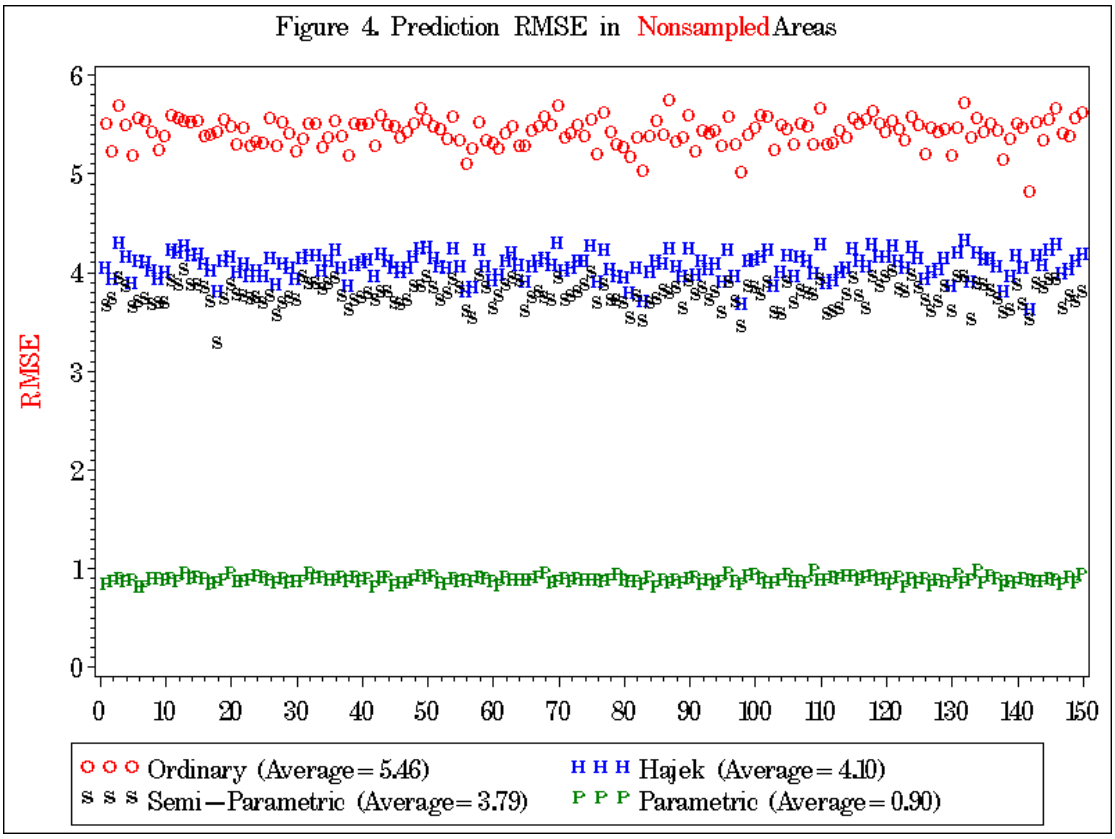


Figure 4. Prediction RMSE in Nonsampled Areas
RMSE
6
5
4
3
2
1
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Ordinary (Average= 5.46)
Hajek (Average= 4.10)
Semi−Parametric (Average= 3.79)
Parametric (Average= 0.90)