SMALL AREA ESTIMATION UNDER INFORMATIVE PROBABILITY SAMPLING OF AREAS AND WITHIN THE SELECTED AREAS

DANNY PFEFFERMANN, MICHAIL SVERCHKOV

ABSTRACT

In this article we show how to predict small area means and obtain valid MSE estimators and confidence intervals when the areas represented in the sample are sampled with unequal probabilities that are possibly related to the true (unknown) area means, and the sampling of units within the selected areas is with probabilities that are possibly related to the outcome values. Ignoring the effects of the sampling process on the distribution of the observed outcomes in such cases may bias the inference very severely. Classical design based inference that uses the randomization distribution of probability weighted estimators cannot be applied for predicting the means of nonsampled areas. We propose simple test statistics for testing the informativeness of the selection of the areas and the sampling of units within the selected areas. The proposed procedures are illustrated by a simulation study and a real application of estimating mean body mass index in counties of the U.S.A, using data from the NHANES III survey.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M07/06

Small Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas

Danny Pfeffermann, Hebrew University and University of Southampton
Michail Sverchkov, Bureau of Labor Statistics and BAE Systems IT

ABSTRACT

In this article we show how to predict small area means and obtain valid MSE estimators and confidence intervals when the areas represented in the sample are sampled with unequal probabilities that are possibly related to the true (unknown) area means, and the sampling of units within the selected areas is with probabilities that are possibly related to the outcome values. Ignoring the effects of the sampling process on the distribution of the observed outcomes in such cases may bias the inference very severely. Classical design based inference that uses the randomization distribution of probability weighted estimators cannot be applied for predicting the means of nonsampled areas. We propose simple test statistics for testing the informativeness of the selection of the areas and the sampling of units within the selected areas. The proposed procedures are illustrated by a simulation study and a real application of estimating mean body mass index in counties of the U.S.A, using data from the NHANES III survey.

Key Words: Body mass index, Bootstrap, Design based inference, Sample distribution, Sample-complement distribution, Sampling weights

Acknowledgement: Opinions expressed in this paper are of the authors and do not constitute a policy of the Bureau of Labor Statistics. The authors thank the referee for very thoughtful comments that improved the article very significantly. The authors thank also Lester Curtin from the National Center of Health Statistics in the U.S. for providing the data used for the empirical application and for his helpful advice.

1. INTRODUCTION

The problem of small area estimation is how to predict the area means or other quantities of interest and assess the prediction errors when the sample sizes in these areas are too small (and possibly zero) to warrant the use of design based methods. It is generally accepted that small area estimation should be based in such cases on statistical models that define ways of borrowing information across areas or over time. See the recent book by Rao (2003) for a comprehensive account of available methods. However, all the models and estimators considered so far assume either that all the areas are represented in the sample or that the sampled areas are selected with equal probabilities. A few studies consider the case where the sampling of units within the selected areas is with unequal selection probabilities that are related to the outcome values, see, Kott (1990), Arora and Lahiri (1997) and Prasad and Rao (1999), but these studies do not consider unit-level observations and only treat the case where the input data consist of direct, design unbiased estimators of the area means. Malec *et al*. (1999) consider unit level observations and use marginal likelihoods and Bayesian methods for inference. We refer to this study in greater detail in Section 10.

In this article we fill this important gap by considering situations where the selection of the areas is with unequal probabilities that are possibly related to the true (unknown) area means, and the sampling of units within the selected areas is with probabilities that are possibly related to the outcome values, even after conditioning on the model covariates. The problem with this kind of sampling designs is that the model holding for the population values no longer holds for the sample data, 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 bias the small area predictors and increase their root mean square errors. For example, the NHANES III survey that is used for the empirical application in Section 10 oversamples minority groups, and if the target variable of interest (body mass index in our application) is related to ethnicity, then clearly any valid inference procedure should account for the sample selection.

In theory, the effect of the sample selection can be controlled by including among the model covariates all the design variables used for the sample selection. However, this is often not practical either because some or all the design variables may not be known or available at the inference stage, or because there are too many of them, making the fitting and validation of such models formidable. Alternatively, one could include in the model the sampling weights as surrogates for the design variables, but this proposition is not operational if the sampling weights are not available for the nonsampled areas or units, which is often the case, particularly in a secondary analysis. Note also that classical design based inference weights the sample observations by the inverse of the sample selection probabilities is not applicable for the prediction of the means in nonsampled areas. This is because design based theory uses the randomization distribution of an estimator over repeated sampling from a fixed finite population as the basis for inference, which can be used for estimating the population quantities of interest, but not for predicting nonsampled values (area means with no samples from these areas in our case).

We use relationships between the 'population distribution', the 'sample distribution' and the 'sample-complement distribution' of an outcome variable developed in Pfeffermann and Sverchkov (1999) and Sverchkov and Pfeffermann (2004), in order to derive approximately unbiased predictors of the means in sampled and nonsampled areas under informative sampling of areas and within the selected areas. We develop estimators for the variances of these predictors and propose simple test statistics for testing the informativeness of the sample selection. The proposed procedures are illustrated by a simulation study and a real application that considers the estimation of mean body mass index (BMI) for counties in the U.S., using NHANES III data.

Section 2 defines the three distributions and shows the relationships between them. Section 3 defines the optimal predictors in sampled and nonsampled areas and Section 4 illustrates the bias resulting from ignoring an informative sampling scheme. In Sections 5 and 6 we establish the theory underlying the proposed prediction procedure, with Section 5 showing step by step how to obtain the predictors of the small area means under a particular model identified for the

sample data and Section 6 developing appropriate variance estimators. Section 7 extends the theory to general sample models. In Section 8 we present the test statistics for testing the informativeness of the sample selection. The simulation results are studied in Section 9, which examines also the performance of confidence intervals for the unknown area means. Section 10 considers the prediction of BMI county means in the U.S. We conclude with a brief summary in Section 11.

2. THE SAMPLE AND SAMPLE-COMPLEMENT DISTRIBUTIONS

Consider a finite population of *N* units belonging to *M* areas, with N_i units in area i. Let *y* define the target variable with value y_{ij} for unit *j* in area i, and denote by x_{ij} the values of corresponding covariates. In what follows we consider the population *y*-values as outcomes of the following two-level random process:

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

We assume a two-stage sampling design 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* and *x*, and possibly also on the population values of design variables *z* used for the sample selection but not included in the model. Denote by I_i and I_{ij} the sample indicator variables for the two sampling stages ($\mathrm{I}_i=1$ iff $i \in s$ and similarly for I_{ij}), and by $w_i = 1/\pi_i$ and $w_{j|i} = 1/\pi_{j|i}$ the first and second stage sampling weights.

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

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

The conditional first level *sample-complement pdf* of u_i, that is, the *pdf* of u_i for area $i \notin s$ is defined in Sverchkov and Pfeffermann (2004) as,

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

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

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

$$\begin{aligned} f_{si}(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1) &\overset{def}{=} f(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1, \mathrm{I}_{ij} = 1) \\ &= \frac{\Pr(\mathrm{I}_{ij} = 1 \mid y_{ij}, x_{ij}, u_i, \mathrm{I}_i = 1) f_p(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1)}{\Pr(\mathrm{I}_{ij} = 1 \mid x_{ij}, u_i, \mathrm{I}_i = 1)}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} f_{ci}(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1) &\overset{def}{=} f(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1, \mathrm{I}_{ij} = 0) \\ &= \frac{\Pr(\mathrm{I}_{ij} = 0 \mid y_{ij}, x_{ij}, u_i, \mathrm{I}_i = 1) f_p(y_{ij} \mid x_{ij}, u_i, \mathrm{I}_i = 1)}{\Pr(\mathrm{I}_{ij} = 0 \mid x_{ij}, u_i, \mathrm{I}_i = 1)}. \end{aligned} \tag{2.4}$$

Here again the *population*, *sample* and *sample-complement pdf*s of y_{ij} are the same if and only if, $\Pr(\mathrm{I}_{ij} = 1 \mid y_{ij}, x_{ij}, u_i, \mathrm{I}_i = 1) = \Pr(\mathrm{I}_{ij} = 1 \mid x_{ij}, u_i, \mathrm{I}_i = 1)\ \forall j$, in which case the sampling of second-level units is *noninformative*. The model defined by (2.1) and (2.3) defines the two-level sample model corresponding to the population model defined by $f_p(u_i)$ and $f_p(y_{ij} \mid x_{ij}, u_i)$; see also Pfeffermann *et al*. (2006).

The following relationships between the population *pdf*, the sample *pdf* and the sample-complement *pdf* are established in Pfeffermann and Sverchkov (1999) and Sverchkov and Pfeffermann (2004) for general pairs of random variables v_1, v_2 measured on elements i of a population P. The symbols E_p, E_s and E_c define respectively the expectations under the three distributions and

$\{\pi_i, w_i\}$ denotes the first order sample inclusion probabilities and the corresponding sampling weights $w_i = 1/\pi_i$.

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

$$E_p(v_{1i}|v_{2i}) = \frac{E_s(w_i v_{1i}|v_{2i})}{E_s(w_i|v_{2i})} \quad ; \quad E_p(\pi_i|v_{2i}) = \frac{1}{E_s(w_i|v_{2i})}, \tag{2.6}$$

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

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

Defining $v_{1i} = u_i$, v_{2i}=*const* yields the relationships holding for the random area effects u_i. Defining $v_{1ij} = y_{ij}$; $v_{2ij} = (x_{ij}, u_i, \mathrm{I}_i = 1)$ and substituting $\pi_{j|i}$ and $w_{j|i}$ for π_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 $D_s = \{(y_{ij}, w_{j|i}, w_i),\ (i,j) \in s;\ x_{kl},\ (k,l) \in U\}$ define the known data. The MSE of a predictor $\hat{\bar{Y}}_i$ with respect to the *population pdf*, given D_s and I_i is,

$$MSE(\hat{\bar{Y}}_i \mid D_s, I_i) = E_p[(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \mid D_s, I_i] = [\hat{\bar{Y}}_i - E_p(\bar{Y}_i \mid D_s, I_i)]^2 + V_p(\bar{Y}_i \mid D_s, I_i). \tag{3.1}$$

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

In what follows we distinguish between *sampled* areas and *nonsampled* areas. For a sampled area i $(I_i = 1)$,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 1) = \frac{1}{N_i} E_p\{[\sum_{l=1}^{N_i} E_p(y_{il} \mid D_s, u_i, \mathrm{I}_{il}, \mathrm{I}_i = 1)] \mid D_s, \mathrm{I}_i = 1\}$$

$$= \frac{1}{N_i}\{\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s]\} \quad (3.2)$$

For area i not in the sample $(I_i = 0)$,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 0) = E_p(\frac{w_i - 1}{E_p[(w_i - 1) \mid D_s, \mathrm{I}_i = 1]} \bar{Y}_i \mid D_s, \mathrm{I}_i = 1) \text{ (by 2.8)}$$

$$= E_p\{\frac{w_i - 1}{E_p[(w_i - 1) \mid D_s, \mathrm{I}_i = 1]}[\frac{1}{N_i}\sum_{k=1}^{N_i} E_p(y_{ik} \mid D_s, u_i, I_i = 1)] \mid D_s, \mathrm{I}_i = 1\}$$

$$= \frac{1}{N_i}\sum_{k=1}^{N_i} E_c[E_p(y_{ik} \mid D_s, u_i, I_i = 1) \mid D_s]. \quad (3.3)$$

4. BIAS OF SMALL AREA PREDICTORS WHEN IGNORING AN INFORMATIVE SAMPLING SCHEME

Consider first a sampled area. Ignoring the sampling scheme of units within the selected area implies an implicit assumption that the *sample-complement* model in the area is the same as the *sample* model such that, $\hat{\bar{Y}}_{i,IGN} = \frac{1}{N_i}\{\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} E_s[E_{si}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s]\}$ (compare with 3.2). Hence,

$$Bias(\hat{\bar{Y}}_{i,IGN}) = E_p[(\hat{\bar{Y}}_{i,IGN} - \bar{Y}_i) \mid D_s, \mathrm{I}_i = 1]$$

$$= \frac{1}{N_i}\sum_{l\notin s_i} E_s[E_{si}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] - \frac{1}{N_i}\sum_{l\notin s_i} E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s]$$

$$= -\frac{1}{N_i} E_s[\sum_{l\notin s_i} \frac{Cov_{si}(y_{il}, w_{lli} \mid D_s, u_i, \mathrm{I}_i = 1)}{E_{si}[(w_{lli} - 1) \mid D_s, u_i, \mathrm{I}_i = 1]} \mid D_s], \quad (4.1)$$

with the last equality following from (2.8). Thus, if the outcomes y_{il} and the sampling weights w_{lli} are correlated given the data in D_s and the random effect u_i, ignoring the sampling scheme yields biased predictors.

Next consider a non-sampled area.

$$Bias(\hat{\bar{Y}}_{i,IGN}) = E_p[(\hat{\bar{Y}}_{i,IGN} - \bar{Y}_i) \mid D_s, \mathrm{I}_i = 0]$$

$$= \frac{1}{N_i}\sum_{k=1}^{N_i} E_s[E_p(y_{ik} \mid D_s, u_i, I_{ik} = 1) \mid D_s] - \frac{1}{N_i}\sum_{k=1}^{N_i} E_c[E_p(y_{ik} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] \quad (4.2)$$

Adding and subtracting $\frac{1}{N_i}\sum_{k=1}^{N_i} E_c[E_p(y_{ik} \mid D_s, u_i, I_{ik} = 1) \mid D_s]$ and use of (2.8) and (2.6) yields,

$$Bias(\hat{\bar{Y}}_{i,IGN}) = -\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{Cov_s(E_p(y_{ik} \mid D_s, u_i, I_{ik}=1), w_i \mid D_s)}{E_s[(w_i - 1) \mid D_s]} - \frac{1}{N_i}E_c[\sum_{k=1}^{N_i}\frac{Cov_{si}(y_{ik}, w_{kli} \mid D_s, u_i, \mathrm{I}_i = 1)}{E_{si}[w_{kli} \mid D_s, u_i, \mathrm{I}_i = 1]} \mid D_s] \quad . \tag{4.3}$$

The first covariance reflects the bias induced by the informative selection of areas. The second covariance reflects the bias induced by the informative sampling process within the selected areas (compare with 4.1). In Section 8 we propose simple test statistics for testing whether the covariances in (4.1) and (4.3) are zero, so that ignoring the sample selection produces unbiased predictors. See also the example below and the empirical results in Section 9.

Example

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

$$y_{ij} = \mu + u_i + e_{ij} \;;\quad u_i \sim N(0, \sigma_u^2)\,,\; e_{ij} \sim N(0, \sigma_e^2)\,, \tag{4.4}$$

with the random effects and residual terms being mutually independent. (In this example $x_{ij} = 1$ for all (i, j)). Consider the common sampling scheme by which *m* areas are sampled with probabilities $\pi_i = c \times N_i$ for some constant *c*, and the second level units are sampled with probabilities $\pi_{j|i} = n_0 / N_i$ (fixed sample size n_0 within the selected areas), such that $\pi_{ij} = \Pr[(i, j) \in s] = \pi_i \pi_{j|i} = const$. Note that the sampling scheme 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. (For example, the areas are school districts, the outcome variable measures children's proficiency; the small districts are the 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,

$$E_p(\pi_i \mid u_i) \prec \exp(Au_i + B + \frac{\sigma_M^2}{2})\,. \tag{4.5}$$

It follows from Pfeffermann *et al.* (1998, example 4.3) that in this case,

$$f_s(u_i) = \frac{E_p(\pi_i | u_i) f_p(u_i)}{E_p(\pi_i)} = N(A\sigma_u^2, \sigma_u^2), \tag{4.6}$$

such that $E_s(u_i) = A\sigma_u^2 < E_p(u_i) = 0$. The fact that the random effects in the sample have a negative expectation is explained by the fact that the sampling scheme undersamples areas with positive random effects. Note, however, that by defining $\mu^* = \mu + A\sigma_u^2$, $u_i^* = u_i - A\sigma_u^2$, the model holding for the outcomes 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 predictor for the mean $\theta_i = \mu + u_i$ of a sampled area ($\mathrm{I}_i = 1$) under the *population model* is also optimal in this case under the *sample model*. (Recall that in this example the sampling scheme within the selected areas is noninformative.)

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

$$f_c(u_i) = \frac{E_p[(1-\pi_i) \mid u_i] f_p(u_i)}{E_p(1-\pi_i)} = \frac{f_p(u_i)}{E_p(1-\pi_i)} - \frac{E_p(\pi_i \mid u_i) f_p(u_i)}{E_p(1-\pi_i)}. \tag{4.7}$$

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 $E_p(\pi_i) = E_p(m)/M$. (For a fixed number m of sampled areas, $E_p(m) = m$.) By (4.7) 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,

$$E_c(u_i) = -\frac{E_p(m) E_s(u_i)}{M - E_p(m)} = -\frac{E_p(m) A\sigma_u^2}{M - E_p(m)} > 0. \tag{4.8}$$

Here again, the positive expectation of the random effects in the nonsampled areas is caused by undersampling of areas with positive random effects. It follows from (4.6) and (4.8) that ignoring the selection scheme of areas and predicting, for example, the means of nonsampled areas by the average of the predictors in the sampled areas yields in this case a prediction bias of,

$$Bias(i \notin s) = A\sigma_u^2 - [-A\sigma_u^2 \frac{E_p(m)}{M - E_p(m)}] = A\sigma_u^2 \frac{M}{M - E_p(m)} < 0. \tag{4.9}$$

Note that the smaller is $E_p(m)$, the larger is the absolute bias.

5. PREDICTION OF SMALL AREA MEANS

In order to facilitate the presentation of our proposed approach, we consider in Sections 5 and 6 a particular sample model and selection scheme. In Section 7 we outline the basic steps in computing the predictors under a general model fitted to the sample data with continuous or discrete outcomes and fixed and random effects, and an arbitrary selection scheme.

The first step of our approach is to fit a model to the sample data, which of course is a necessary step in any small area estimation application. Note that although we consider informative sampling, the sample model can be identified and estimated from the sample data using standard techniques, see Rao (2003) for small area model identification and diagnostic methods. In this and the next section we suppose that the sample model is the 'nested error regression model',

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

Suppose that the sampled areas are selected with probabilities, π_i, $i = 1...m$ and that n_i units are sampled from selected area i with probabilities $\pi_{j|i}$, for which the sampling weights, $w_{j|i} = (1/\pi_{j|i})$ satisfy,

$$E_{si}(w_{j|i} \mid x_{ij}, y_{ij}, u_i, \mathrm{I}_i = 1) = E_{si}(w_{j|i} \mid x_{ij}, y_{ij}, \mathrm{I}_i = 1) = k_i \exp(a x_{ij} + b y_{ij}) \, , \tag{5.2}$$

where $k_i = N_i (n_i)^{-1} \sum_{j=1}^{N_i} \exp(-a x_{ij} - b y_{ij}) / N_i$ (follows from (2.6)), and a and b are fixed (possibly unknown) constants. (If x_{ij} is a vector, a is a vector. If $x_{i0} = constant$, we assume $a_0 = 1$ for uniqueness). Note that for large areas, (large N_i), $\sum_{j=1}^{N_i} \exp(-a x_{ij} - b y_{ij}) / N_i \cong E_p[\sum_{j=1}^{N_i} \exp(-a x_{ij} - b y_{ij}) / N_i] = const$, such that $k_i \cong (N_i / n_i) \times const$. As becomes evident below, for sufficiently small sampling fractions the predictors for sampled and nonsampled areas do not depend on a and k_i.

Remark 1: It follows from Pfeffermann *et al.* (1998) that under the sampling scheme (5.2) the population model is also of the form (5.1), but with different parameters, if the areas are selected with probabilities π_i satisfying, $E(\pi_i \mid \theta_i) \prec \exp[\gamma_0 \theta_i + z_i'\gamma]$, where $\theta_i = \bar{X}_i \beta + u_i$ are the area means, z_i represents area level design variables and (γ_0, γ) are fixed coefficients. The

model (5.1) is in common use for small area estimation under noninformative sampling (in which case the population and sample models coincide), see, e.g., Battese *et al*. (1988).

In what follows we only assume knowledge of the form of the sample model (5.1) and the conditional expectations in (5.2), but not the form of the population model or the relationship between the area selection probabilities and the area means.

Remark 2: As with the sample model (5.1), the expectation in (5.2) refers to the sample distribution within the sampled areas. The relationship in the sample between the sampling weights and the outcome values can be identified and estimated therefore from the sample data, see Skinner (1994) and 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 paper we do not model this relationship. See Pfeffermann *et al*. (2006) 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 his article assumes implicitly that all the population areas are sampled.

The analysis that follows assumes known model parameters. In practice, the unknown model parameters are replaced under the frequentist approach by sample estimates, yielding the corresponding 'empirical predictors'. Maximum likelihood estimation of the model parameters has to be based in the present case on the sample distribution of the sample outcomes (the distribution obtained from (2.1) and (2.3) that conditions on the selected units and areas), as identified from the sample data. Alternatively, the unknown model parameters can possibly be estimated by application of the 'method of moments', depending on the underlying model. See the empirical study in Sections 9 and 10.

We make the following mild assumption:

Ass.1- $f_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) = f_{ci}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)$, implying that observed and unobserved outcomes in a sampled area are independent when conditioning on

the area random effect. To see that this is a mild assumption note that if the population outcomes are independent given the random effect, then by (2.7),

$$f_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) = \frac{E_p[(1-\pi_{il}) \mid y_{il}, D_s, u_i, \mathrm{I}_i = 1]}{E_p[(1-\pi_{il}) \mid D_s, u_i, \mathrm{I}_i = 1]} f_p(y_{il} \mid x_{ij}, u_i, \mathrm{I}_i = 1).$$

Furthermore, by (5.2), and the fact that $k_i \cong (N_i / n_i) \times const$ (assuming large N_i), the expectation in the numerator likewise only depends on y_{il} and not on the observations y_{ij}, $j \in s_i$.

As established in Section 3, the optimal predictor for a sampled area i is, $E_p(\overline{Y}_i \mid D_s, \mathrm{I}_i = 1) = [\sum_{j \in s_i} y_{ij} + \sum_{l \notin s_i} E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s]] / N_i$. In order to compute the expectations $E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1)$ we proceed as follows: First, by (2.7), (5.1), (5.2),

$$f_{ci}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) = \frac{[E_{si}(w_{l\mid i} \mid x_{il}, y_{il}, u_i, \mathrm{I}_i = 1) - 1] f_{si}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)}{E_{si}(w_{l\mid i} \mid x_{il}, u_i, \mathrm{I}_i = 1) - 1}$$

$$= \frac{\lambda_{il}}{\lambda_{il} - 1} \frac{1}{\sigma_e} \phi\left(\frac{y_{ij} - u_{il} - b\sigma_e^2}{\sigma_e}\right) - \frac{1}{\lambda_{il} - 1} \frac{1}{\sigma_e} \phi\left(\frac{y_{il} - u_{il}}{\sigma_e}\right), \qquad (5.3)$$

where $u_{il} = x_{il}'\beta + u_i$, $\lambda_{il} = k_i \exp[(b^2\sigma_e^2 / 2) + a x_{il} + b u_{il}] = E_{si}(w_{l\mid i} \mid x_{il}, u_i, \mathrm{I}_i = 1)$ and ϕ is the standard normal *pdf*. Note that if $b = 0$ (the selection probabilities within the sampled areas only depend on the *x*-values so that the sampling is noninformative), the *pdf* in (5.3) reduces to the sample normal density (5.1).
By (5.3),

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

and hence by *Ass.*1,

$$E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] = E_s[E_{ci}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) \mid D_s]$$
$$= E_s[(u_{il} + \frac{\lambda_{il}}{\lambda_{il} - 1} b\sigma_e^2) \mid D_s] \quad . \qquad (5.5)$$

The last expectation in (5.5) is with respect to the sample distribution of $u_i \mid D_s, \mathrm{I}_i = 1$. Under the sample model (5.1), this distribution is normal with mean $\hat{u}_i = \gamma_i[\overline{y}_i - \overline{x}_i'\beta]$ and variance $\sigma_i^2\gamma_i$, where $(\overline{y}_i, \overline{x}_i) = \sum_{j=1}^{n_i} (y_{ij}, x_{ij}) / n_i$ are the

sample means of (*y,x*) in sampled area i, $\gamma_i = \sigma_u^2/[\sigma_u^2 + \sigma_i^2]$ and $\sigma_i^2 = \sigma_e^2/n_i = Var_s(\bar{y}_i \mid u_i)$. Thus, for a sampled area the expectation $E_{ci}(y_{il} \mid D_s, \mathrm{I}_i = 1)$ is obtained by computing the expectation in the right hand side of (5.5) with respect to the normal distribution of $u_i \mid D_s, \mathrm{I}_i = 1$ defined above. We find that,

$$E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] = (x_{il}'\beta + \hat{u}_i) + b\sigma_e^2 E_s[(1-\lambda_{il}^{-1})^{-1} \mid D_s]. \quad (5.6)$$

Note that if *b*=0 (noninformative sampling within the area), $E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] = x_{il}'\beta + \hat{u}_i$, which is the standard result.

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

$$E_s[(1-\lambda_{il}^{-1})^{-1} \mid D_s] \cong 1, \quad (5.7)$$

in which case by (5.6), $E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s] \cong (x_{il}'\beta + \hat{u}_i) + b\sigma_e^2$, where $\hat{u}_{il} = x_{il}'\beta + \hat{u}_i$.

It follows from (3.2), (5.6) and (5.7) that for given parameters $\{\beta, b, \sigma_u^2, \sigma_e^2\}$, the mean $\bar{Y}_i$ of sampled area i can be predicted as,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 1) = \frac{1}{N_i}\{(N_i - n_i)\hat{\theta}_i + n_i[\bar{y}_i + (\bar{X}_i - \bar{x}_i)'\beta] + (N_i - n_i)b\sigma_e^2\}, \quad (5.8)$$

where $\hat{\theta}_i = \hat{u}_i + \bar{X}_i\beta$ is the optimal predictor of the sample model mean $\theta_i = \bar{X}_i'\beta + u_i \; = E_{si}(\bar{Y}_i \mid u_i)$. The last term in (5.8) corrects for the sample selection effects, that is, the difference between the sample-complement expectation and the sample expectation in sampled areas. Note again that under noninformative sampling (*b*=0), the predictor (5.8) reduces to the optimal predictor under noninformative (Rao, 2003, Eq. 7.2.37).

The optimal predictor for *nonsampled* areas is defined in (3.3) to be,

$$E_p(\bar{Y}_i \mid D_s, \mathrm{I}_i = 0)$$
$$= \frac{1}{N_i}\sum_{k=1}^{N_i} E_c[E_p(y_{ik} \mid D_s, u_i.I_i = 1) \mid D_s] = \frac{1}{N_i}\sum_{k=1}^{N_i} E_c[E_p(y_{ik} \mid x_{ik}, u_i.I_i = 1) \mid D_s].$$ (The last equality follows from the fact that the outcomes y_{ik} refer to a nonsampled area. See also *Ass.*2 below.) By (2.8) and then (2.6),

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

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

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

Estimating the two sample expectations in the right hand side of (5.10) by the corresponding sample means and substituting $\hat{u}_i = \gamma_i[\bar{y}_i - \bar{x}_i'\beta]$ for u_i yields the following estimate for $E_{ik} = E_c[E_p(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1) \mid D_s]$,

$$E_{ik} = x_{ik}'\beta + b\sigma_e^2 + \frac{\sum_{i \in s}(w_i - 1)\hat{u}_i}{\sum_{i \in s}(w_i - 1)}. \quad (5.11)$$

It follows from (3.3) and (5.11) that for given parameters $\{\beta, b, \sigma_u^2, \sigma_e^2\}$, the mean $\bar{Y}_i$ of area i not in the sample can be predicted as,

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

The term $\sum_{i \in s}(w_i - 1)\hat{u}_i / \sum_{i \in s}(w_i - 1)$ corrects for the fact that the mean of the random effects for areas outside the sample is different from zero under informative sampling of the areas.

6. MSE ESTIMATION

Estimating $MSE(\hat{\bar{Y}}_i \mid D_s, I_i) = E_p[(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \mid D_s, I_i)]$ for the predictors considered in section 5 requires strict model assumptions that could be hard to validate. This

is largely due to the conditioning on the design information D_s. In order to deal with this problem, we estimate instead $MSE(\hat{\bar{Y}}_i \mid X, \mathrm{I}_i) = E_p[MSE(\hat{\bar{Y}}_i \mid D_s, \mathrm{I}_i) \mid X, \mathrm{I}_i]$, where $X = \{x_{ij}, (i,j) \in U\}$. Note that $MSE(\hat{\bar{Y}}_i \mid D_s, \mathrm{I}_i)$ can be viewed as random, such that $MSE(\hat{\bar{Y}}_i \mid X, \mathrm{I}_i)$ defines its 'best predictor' with respect to the mean square loss function under the distribution $f_{D_s \mid X, \mathrm{I}_i}$.

Denote by $\hat{\bar{Y}}_i$ the predictor defined by (5.8) if $i \in s$ or by (5.12) if $i \notin s$. For what follows we make the following additional mild assumptions:

*Ass.*2 $Co\mathrm{v}[y_{ij}, y_{mk} \mid \mathrm{I}_i = 1, \mathrm{I}_m = 0] = 0$; $Co\mathrm{v}[y_{ij}, y_{ik} \mid u_i, \mathrm{I}_i = 1, \mathrm{I}_{ij} = \mathrm{I}_{ik} = 0] = 0$; implying that observations in sampled areas are uncorrelated with observations in nonsampled areas, and that the unobserved outcomes in a sampled area are uncorrelated, conditionally on the realization of the random effect. The first assumption will always hold if the random effects are independent between the areas. The second assumption is also not restrictive if the population observations are conditionally independent, because by extending Remark 2 of Sverchkov and Pfeffermann (2004) to the case of a joint distribution for a pair of units, it follows that for small sampling fractions (the common situation in small area estimation), the joint sample-complement distribution and the corresponding population distribution are approximately the same.

*Ass.*3 $Co\mathrm{v}[y_{ij}, y_{ik} \mid u_i, \mathrm{I}_i = 0] = 0$; implying that the outcomes in a nonsampled area are uncorrelated conditional on the realization of the random effect. This assumption holds as long as the selection of the areas only depends on the area means.

*Ass.*4 The predictor $\hat{\bar{Y}}_i$, $i \notin s$ is approximately unbiased for $E_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0)$ in the sense that, $E_s[E_{si}(\hat{\bar{Y}}_i \mid X, u_i, \mathrm{I}_i = 1) \mid X] \cong E_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0)$ (follows from Section 5).

Consider first *sampled* areas. Denote $Y_{Ri} = Y_i - \sum_{j \in s_i} y_{ij}$ where $Y_i = N_i \bar{Y}_i$, such that $\hat{Y}_{Ri} = N_i \hat{\bar{Y}}_i - \sum_{j \in s_i} y_{ij}$. Noting that $X \subset D_s$ and that conditional on $(D_s, u_i, \mathrm{I}_i = 1)$, $\hat{Y}_{Ri} - E_c[Y_{Ri} \mid D_s, u_i, \mathrm{I}_i = 1]$ is constant, it follows from *Ass.*2 that,

$$E_p[(\hat{Y}_i - Y_i)^2 \mid X, \mathrm{I}_i = 1] = E_p\{E_p[(\hat{Y}_i - Y_i)^2 \mid D_s, u_i, \mathrm{I}_i = 1] \mid X, \mathrm{I}_i = 1\}$$

$$= E_p\{[\hat{Y}_{Ri} - E_{ci}(Y_{Ri} \mid D_s, u_i, \mathrm{I}_i = 1)]^2 \mid X, \mathrm{I}_i = 1\}$$

$$+ E_p\{E_p[(Y_{Ri} - E_{ci}(Y_{Ri} \mid D_s, u_i, \mathrm{I}_i = 1))^2 \mid D_s, u_i, \mathrm{I}_i = 1] \mid X, \mathrm{I}_i = 1\}$$

$$= E_p\{[\hat{Y}_{Ri} - E_{ci}(Y_{Ri} \mid D_s, u_i, \mathrm{I}_i = 1)]^2 + E_{ci}[(Y_{Ri} - E_{ci}(Y_{Ri} \mid X, u_i, \mathrm{I}_i = 1))^2 \mid X, u_i, \mathrm{I}_i = 1] \mid X, \mathrm{I}_i = 1\}$$

$$= E_s[E_{si}(F(u_i, D_s) \mid X, u_i, \mathrm{I}_i = 1) \mid X], \tag{6.2}$$

where $F(u_i, D_s) = [\hat{Y}_{Ri} - E_c(Y_{Ri} \mid D_s, u_i, \mathrm{I}_i = 1)]^2 + E_c[(Y_{Ri} - E_c(Y_{Ri} \mid X, u_i, \mathrm{I}_i = 1))^2 \mid X, u_i, \mathrm{I}_i = 1]$.

By (5.3) and (5.4),

$$E_{ci}(y_{il}^2 \mid x_{il}, u_i, \mathrm{I}_i = 1) = \frac{\lambda_{il}}{\lambda_{il} - 1}[\sigma_e^2 + (u_{il} + b\sigma_e^2)^2] - \frac{1}{\lambda_{il} - 1}[\sigma_e^2 + u_{il}^2],$$

$E_{ci}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) = u_{il} + \frac{\lambda_{il}}{\lambda_{il} - 1} b\sigma_e^2$; $u_{il} = x_{il}'\beta + u_i$. Hence,

$E_{ci}[(y_{il} - E_{ci}[y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1])^2 \mid x_{il}, u_i, \mathrm{I}_i = 1] = \sigma_e^2 - \frac{\lambda_{il} b^2 \sigma_e^4}{(\lambda_{il} - 1)^2}$. Note that under (5.2) the last (sample-complement) variance is smaller than $\sigma_e^2 = Var_s(e_{ij} \mid \mathrm{I}_{ij} = 1)$ in (5.1), unless the sampling within the areas is noninformative ($b = 0$). Thus, by *Ass.*1 and *Ass.*2 $F(u_i, D_s)$ in (6.2) can be written as,

$$F(u_i, D_s) = [\hat{Y}_{Ri} - \sum_{l \notin s_i}(u_{il} + \frac{\lambda_{il}}{\lambda_{il} - 1} b\sigma_e^2)]^2 + \sum_{l \notin s_i}(\sigma_\varepsilon^2 - \frac{\lambda_{il} b^2 \sigma_\varepsilon^4}{(\lambda_{il} - 1)^2}). \tag{6.3}$$

All the terms in (6.3) are either fixed values or that they are functions of the sample data in D_s and the random effect u_i in sampled area i. It follows therefore that $MSE(\hat{\bar{Y}}_i \mid X, I_i = 1) = (1/N_i^2) E_s[E_{si}(F(u_i, D_s) \mid X, u_i, \mathrm{I}_i = 1) \mid X]$ can be estimated by the following parametric bootstrap procedure (see Remark 4 below):

1. Estimate $a, b, k_i, \beta, \sigma_u^2, \sigma_e^2$ (see Section 9),

2. Generate *B* bootstrap samples $\{u_i^b, y_{ij}^b\}$, $i = 1,...,m$, $j = 1,...,n_i$ from the sample model (5.1) with parameters $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$, using the covariates $x_{ij}, (i, j) \in s$. Compute $w_{jli}^b = \hat{k}_i \exp(\hat{a} x_{ij} + \hat{b} y_{ij}^b)$.

3. Re-compute the predictors $\hat{Y}_{Ri}^b \quad i = 1...m, b = 1,...,B$ (with new parameter estimates) and compute $F^b(u_i^b, D_s^b)$ defined by (6.3); the new parameter estimates are only used for the computation of $\hat{Y}_{Ri}^b$, the other terms of $F^b(u_i^b, D_s^b)$ are computed using the original parameter estimates.

4. Estimate,

$$\hat{MSE}(\hat{\bar{Y}}_i \mid X, I_i = 1) = \frac{1}{B}\sum_{b=1}^{B} F^b(u_i^b, D_s^b) / N_i^2 \qquad (6.4)$$

Remark 3: The bootstrap estimator (6.4) ignores the contribution to the variance from estimating the hyper-parameters $\{\beta, k_i, a, b, \sigma_u^2, \sigma_e^2\}$. Accounting for this extra source of variation requires a 'double bootstrap' procedure. See Hall and Maiti (2006) for bootstrap bias corrections in small area estimation that warrant MSE estimation of order $O(1/m^2)$.

Next consider *nonsampled* areas. By Ass.2 and *Ass.*4,

$$E_p\{(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \mid X, \mathrm{I}_i = 0\}$$

$$= E_p\{[\hat{\bar{Y}}_i - E_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0)]^2 + E_p[\bar{Y}_i - E_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0)]^2 \mid X, \mathrm{I}_i = 0\}$$

$$\cong E_s\{\{E_{si}[\hat{\bar{Y}}_i - E_s(E_{si}(\hat{\bar{Y}}_i \mid X, u_i, \mathrm{I}_i = 1) \mid X)]^2 \mid X, u_i, \mathrm{I}_i = 1\} \mid X\} + Var_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0).$$

$$= E_s[E_{si}(G(u_i, D_s) \mid X, u_i, \mathrm{I}_i{=}1) | X] + Var_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0), \qquad (6.5)$$

where $G(u_i, D_s) = [\hat{\bar{Y}}_i - E_s(E_{si}(\hat{\bar{Y}}_i \mid X, u_i, \mathrm{I}_i = 1) \mid X)]^2$. The first expression in (6.5) can be estimated similarly to the estimation of $MSE(\hat{\bar{Y}}_i \mid X, I_i = 1)$ above, that is, by applying the first 3 steps of the bootstrap procedure to obtain realizations $\hat{\bar{Y}}_i^b$ for nonsampled area i, and then estimating,

$$E_s[E_{si}(G(u_i, D_s) \mid X, u_i, \mathrm{I}_i{=}1) | X] = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_i^b - \hat{\bar{Y}}_{i,A})^2 \; ; \quad \hat{\bar{Y}}_{i,A} = \frac{1}{B}\sum_{b=1}^{B}\hat{\bar{Y}}_i^b . \qquad (6.6)$$

Note that like with $F(u_i, D_s)$ in (6.3), $\hat{\bar{Y}}_i$ only uses the sample data and hence it is only needed to generate data from the sample model.

In order to estimate the second variance in (6.5) we use the following variance decomposition,

$$\begin{aligned} Var_p(\bar{Y}_i \mid X, \mathrm{I}_i = 0) = {} & Var_p[E_p(\bar{Y}_i \mid u_i, X, \mathrm{I}_i = 0) \mid X, \mathrm{I}_i = 0] \\ & + E_p[Var_p(\bar{Y}_i \mid u_i, X, \mathrm{I}_i = 0) \mid X, \mathrm{I}_i = 0] \end{aligned} . \qquad (6.7)$$

Under *Ass.*3, the second component in (6.7) is simply,

$$E_p[Var_p(\bar{Y}_i \mid u_i, X, I_i = 0) \mid X, I_i = 0] = \sigma_e^2 / N_i . \qquad (6.8)$$

This result follows from the fact that under the sample model (5.1) and the

sampling scheme (5.2), $E_p(e_{ij} \mid u_i, X, I_i = 0) = const$ and $Var_p(y_{ij} \mid u_i, x_{ij})$ $= Var_{si}(y_{ij} \mid u_i, x_{ij}) = \sigma_e^2$ (Pfeffermann *et al.* 1998).

Next consider the first term of (6.7). By similar arguments,

$$Var_p[E_p(\bar{Y}_i \mid u_i, X, \mathrm{I}_i = 0) \mid X, \mathrm{I}_i = 0] = Var_p(u_i \mid X, \mathrm{I}_i = 0) \tag{6.9}$$

Let $\tilde{\xi} = \{\tilde{u}_i, \tilde{e}_{ij}, \tilde{\mathrm{I}}_{ij}, \tilde{\mathrm{I}}_i, \tilde{w}_i, (i,j) \in U\}$ be a generic random vector distributed identically but independently of $\xi = \{u_i, e_{ij}, \mathrm{I}_{ij}, \mathrm{I}_i, w_i, (i,j) \in U\}$ given X under the population distribution, that is, $P_p(\tilde{\xi} \in A \mid X) = P_p(\xi \in A \mid X)$, for every set A belonging to the σ-algebra generated by ξ. Define, $\tilde{y}_{ij} = x_{ij}\beta + \tilde{u}_i + \tilde{e}_{ij}$. Then by

*Ass.*3, $Var_p[\sum_{i \in s} \frac{1}{n_i} \sum_{j \in s_i} (\tilde{y}_{ij} - x'_{ij}\beta) \mid X, \tilde{\mathrm{I}}_i = 0]$

$= Var_p[(\sum_{i \in s} \tilde{u}_i + \sum_{i \in s} \frac{1}{n_i} \sum_{j \in s_i} \tilde{e}_{ij}) \mid X, \tilde{\mathrm{I}}_i = 0] = m Var_p(u_i \mid X, \mathrm{I}_i = 0) + \sum_{i \in s} \frac{\sigma_e^2}{n_i}$, such that

$$Var_p(u_i \mid X, \mathrm{I}_i = 0) = \frac{1}{m} Var_p[\sum_{i \in s} \frac{1}{n_i} \sum_{j \in s_i} (\tilde{y}_{ij} - x'_{ij}\beta) \mid \tilde{\mathrm{I}}_i = 0] - \frac{1}{m} \sum_{i \in s} \frac{\sigma_e^2}{n_i}. \tag{6.10}$$

Let $\tilde{r}_i = \frac{1}{n_i} \sum_{j \in s_i} (\tilde{y}_{ij} - x_{ij}\,'\beta)$; $r_i = \frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - x_{ij}\,'\beta)$. Then, by (6.10) and (2.8),

$$Var_p\left(u_i \mid X, \mathrm{I}_i = 0\right) = \frac{1}{m} \sum_{i \in s} E_s\{\frac{\tilde{w}_i - 1}{E_s(\tilde{w}_i - 1)}[\tilde{r}_i - E_s \frac{\tilde{w}_i - 1}{E_s(\tilde{w}_i - 1)} \tilde{r}_i]^2\} - \frac{1}{m} \sum_{i \in s} \frac{\sigma_e^2}{n_i}. \tag{6.11}$$

It follows from (6.8) and (6.11), that the variance for nonsampled areas can be estimated as,

$$\hat{Var}_p(\bar{Y}_{ci} \mid X, \mathrm{I}_i = 0) = \sum_{i \in s} \frac{w_i - 1}{\sum_{i \in s}(w_i - 1)} [\hat{r}_i - \sum_{i \in s} \frac{w_i - 1}{\sum_{i \in s}(w_i - 1)} \hat{r}_i]^2 + \frac{\hat{\sigma}_e^2}{N_i} - \frac{1}{m} \sum_{i \in s} \frac{\hat{\sigma}_e^2}{n_i}, \tag{6.12}$$

where $\hat{\beta}, \hat{\sigma}_e^2$ are sample estimates of β, σ_e^2 and $\hat{r}_i = \frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - x_{ij}\,'\hat{\beta})$. Note that the expectations under the sample distribution in (6.11) are replaced in (6.12) by the corresponding sample means.

7. PREDICTION OF THE SMALL AREA MEANS UNDER A GENERAL SAMPLE MODEL

In Section 5 we consider a particular sample model and selection scheme as defined by (5.1) and (5.2). Below we outline the basic steps in computing the predictors under a general model fitted to the sample data, with continuous or

discrete outcomes and fixed and random effects. We assume informative sampling of areas and within the areas and maintain *Ass.*1 of Section 5.

As implied by (3.2)-(3.3), computation of the predictors of the small area means requires estimating $E_s[E_{ci}(y_{il} \mid D_s, u_i, \mathrm{I}_i = 1) \mid D_s]$ for sampled areas and $E_c[E_p(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1) \mid D_s]$ for the nonsampled areas. Consider first sampled areas. By *Ass.*1 and then (2.7),

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

$$= E_s[\int y_{il} f_{ci}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) dy_{il} \mid D_s]$$

$$= E_s[\int y_{il} \frac{E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, u_i, \mathrm{I}_i = 1) - 1}{E_{si}(w_{l\mathrm{I}i} \mid x_{il}, u_i, \mathrm{I}_i = 1) - 1} f_{si}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) dy_{il} \mid D_s]$$

$$= \int \frac{\int y_{il} \{E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, u_i, \mathrm{I}_i = 1) - 1\} f_{si}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) dy_{il}}{E_{si}(w_{l\mathrm{I}i} \mid x_{il}, u_i, \mathrm{I}_i = 1) - 1} f_s(u_i \mid D_s) du_i . \qquad (7.1)$$

The last expression is a function of the sample models $f_{si}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1)$, $E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, u_i, \mathrm{I}_i = 1)$ and $f_s(u_i \mid D_s, \mathrm{I}_i = 1)$, all of which can be identified from the sample data, see Remarks 1 and 2 in Section 5. On the other hand, if the distribution $f_s(u_i \mid D_s, \mathrm{I}_i = 1)$ cannot be identified properly, one can estimate (7.1) by the sample mean $\bar{H}_s = \sum_{i \in s} H_{x_{il}}(u_i)/n$, where,

$$H_{x_{il}}(u_i) = \int y_{il} \frac{E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, u_i, \mathrm{I}_i = 1) - 1}{E_{si}(w_{l\mathrm{I}i} \mid x_{il}, u_i, \mathrm{I}_i = 1) - 1} f_{si}(y_{il} \mid x_{il}, u_i, \mathrm{I}_i = 1) dy_{il} \qquad (7.2)$$

In practice it would often be sensible to assume $E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, u_i, \mathrm{I}_i = 1) = E_{si}(w_{l\mathrm{I}i} \mid y_{il}, x_{il}, \mathrm{I}_i = 1)$ like in (5.2). Once the sample models are identified and the integrals in (7.1) are computed, either analytically or numerically if necessary, one can proceed similarly to Section 5, replacing unknown parameters by sample estimates.

Consider next nonsampled areas. By (3.3), estimating the means of such areas requires estimating, $E_c[E_p(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1) \mid D_s]$. Similar arguments to (5.9) imply,

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

$$= \frac{E_s\{(w_i - 1)\dfrac{E_{si}[E_{si}(w_{k|i} \mid y_{ik}, x_{ik}, u_i, \mathrm{I}_i = 1) y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1]}{E_{si}[E_{si}(w_{k|i} \mid y_{ik}, x_{ik}, u_i, \mathrm{I}_i = 1) \mid x_{ik}, u_i, \mathrm{I}_i = 1]} \mid D_s\}}{E_s(w_i \mid D_s) - 1}$$

$$= \frac{E_s[(w_i - 1)\dfrac{\int E_{si}\{w_{k|i} \mid y_{ik}, x_{ik}, u_i, \mathrm{I}_i = 1\} y_{ik} f_{si}(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1) dy_{ik}}{\int E_{si}\{w_{k|i} \mid y_{ik}, x_{ik}, u_i, \mathrm{I}_i = 1\} f_{si}(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1) dy_{ik}} \mid D_s]}{E_s(w_i \mid D_s) - 1}$$

$$\approx \frac{\sum_{r \in s}(w_r - 1) K_{x_{ik}}(u_r)}{\sum_{r \in s}(w_r - 1)}, \qquad (7.3)$$

where $K_{x_{ik}}(u_r)$ is the ratio of the two integrals. Note that $K_{x_{ik}}(u_r)$ is again a function of the sample models $f_{si}(y_{ik} \mid x_{ik}, u_i, \mathrm{I}_i = 1)$ and $E_{si}(w_{k|i} \mid y_{ik}, x_{ik}, u_i, , \mathrm{I}_i = 1)$, as when predicting the means of sampled areas, and hence the prediction of means of nonsampled areas follows the same steps as outlined below (7.1).

8. TESTING FOR PREDICTION BIAS

Evidently, predicting the small area means under informative sampling is more complicated than under noninformative sampling. Also, the predictors developed for the case of informative sampling generally have larger variances than the variances of the optimal predictors under a given population model, if the sampling process is not informative. Thus, it is important to test the informativeness of the sample selection and if found noninformative, use standard optimal procedures. In what follows we propose simple test statistics for testing whether ignoring the sample selection biases the predictors. We study the performance of these tests in the simulation study described in Section 9.

8.1 *Testing whether ignoring the selection of areas biases the predictors.*

As implied by (4.3), the selection of areas does not bias the optimal predictors under noninformative sampling if, $\dfrac{Cov_s\{[\frac{1}{N_i}\sum_{k=1}^{N_i} E_p(y_{ik} \mid D_s, u_i, I_{ik} = 1)], w_i \mid D_s\}}{E_s[(w_i - 1) \mid D_s]} = 0$.

However, we only need to test $Corr_s(u_i, w_i) = 0$ because if the true area means are functions also of the covariate means $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij} / N_i$ as, for example, under the model (5.1), dependence of w_i on $\bar{X}_i$ alone does not bias the predictors. To see this, note that the sample *pdf* of the area mean θ_i is by (2.5) and (2.6), $f_s(\theta_i \mid \bar{X}_i) = E_s(w_i \mid \bar{X}_i) f_p(\theta_i \mid \bar{X}_i) / E_s(w_i \mid \theta_i, \bar{X}_i)$, and if $E_s(w_i \mid \theta_i, \bar{X}_i) = E_s(w_i \mid \bar{X}_i)$, $f_s(\theta_i \mid \bar{X}_i) = f_p(\theta_i \mid \bar{X}_i)$. This is true for general population models.

For testing $H_0 : Corr_s(w_i, u_i) = 0$ we would ideally regress w_i against u_i but the random effects are unobservable. Thus, we regress instead w_i against the estimates $\hat{u}_i$ as computed under the sample model. For the model (5.1), the estimates are defined in Section 5 as, $\hat{u}_i = \hat{\gamma}_i[\bar{y}_i - \bar{x}_i'\beta]$. Writing $\hat{u}_i = u_i + \eta_i$, it is clear that $Cov(w_i, \eta_i) = 0$ such that testing H_0 can be implemented by regressing $w_i = \delta_0 + \delta \hat{u}_i + \varsigma_i$ and testing $H_0 : \delta = 0$, using the conventional t-statistic,

$$t^A = \hat{\delta}_{OLS} / \sqrt{\hat{Var}(\hat{\delta}_{OLS})} \tag{8.1}$$

Under H_0 and some mild conditions, t^A has approximately a t-distribution with $(m-2)$ degrees of freedom, irrespective of the underlying sample model. The hypothesis (8.1) refers to the sample distribution, thus justifying estimating δ by OLS. The drawback of the test statistic (8.1) is that it may not be very powerful if $Var(\eta_i)$ is large. An alternative test can possibly be constructed by noting that $\hat{u}_i = u_i + \eta_i$ and using errors in variables techniques.

8.2 *Testing whether ignoring the sampling schemes within the selected areas biases the predictors.*

By (4.1) and (4.3), sampling within the areas does not bias the optimal predictors under noninformative sampling if, $Cov_{si}(y_{il}, w_{l|i} \mid D_s, u_i, \mathrm{I}_i = 1) = 0$ for $l \notin s_i$. Assuming that, $Cov_{si}(y_{il}, w_{l|i} \mid D_s, u_i, \mathrm{I}_i = 1) = Cov_{si}(y_{il}, w_{l|i} \mid x_{il}, \mathrm{I}_i = 1)$, the ignorability of the sample selection within the selected areas can be tested by testing $Corr_{si}(w_{j|i}, y_{ij} \mid x_{ij}, \mathrm{I}_i = 1) = 0$. This can be implemented by regressing $w_{j|i} = \gamma_{0i} + \gamma_{1i} x_{ij} + \gamma_{2i} y_{ij} + \eta_{ij}$ and testing $H_0 : \gamma_{2i} = 0$ for every $i \in s$. However, with

a large number of sampled areas, testing H_0 for every area is not practical, and with small sample sizes within the areas, the tests have low power. Assuming the same sampling scheme within each of the areas, a more powerful and practical test statistic is therefore,

$$F^{w}_{max} = max[F_i, i = 1...m] \qquad (8.2)$$

where $F_i = [\hat{\gamma}_{2i} / S\hat{D}(\hat{\gamma}_{2i})]^2$ and $\hat{\gamma}_{2i}, S\hat{D}(\hat{\gamma}_{2i})$ are respectively the OLS estimator and its estimated standard deviation. Under the null hypothesis $H_0 : \gamma_{2i} = 0$, $F_i \sim F(1, n_i - 3)$. Computation of the percentiles of F^{w}_{max} for given sampled areas and sample sizes is straightforward.

Remark 4: Instead of testing $Corr_{si}(w_{j|i}, y_{ij} \mid x_{ij}, \mathrm{I}_i = 1) = 0$ by fitting a linear model, one can test whether $E_{si}(w_{j|i} \mid y_{ij}, x_{ij}, \mathrm{I}_i = 1) = E_{si}(w_{j|i} \mid x_{ij}, \mathrm{I}_i = 1)$ allowing for other relationships between the weights $w_{j|i}$ and (x_{ij}, y_{ij}), like the relationship (5.2), if such relationships can be surmised or identified from the sampled data. Note from (2.5) and (2.6) that $E_{si}(w_{j|i} \mid y_{ij}, x_{ij}, \mathrm{I}_i = 1) = E_{si}(w_{j|i} \mid x_{ij}, \mathrm{I}_i = 1)$ implies that the population and sample distributions within the selected areas are the same.

9. MONTE-CARLO SIMULATION STUDY

In order to illustrate the biases that can occur when ignoring an informative sampling scheme and to assess the performance of the procedures developed in this article, we designed a small simulation study. The study consists of the following steps:

1- Generate area indexes $i = 1,..., M = 50$ and population sizes, $N_i = Int\{1000 \times (0.5 + \xi_i)\}$; $\xi_i \sim U[0,1]$. Generate auxiliary values $x_{ij} = (50, t_{ij})'$, $t_{ij} = 1 + 3 \times Int[i - \frac{50}{3} \times Int(\frac{3}{50} \times i)]/10 + \varsigma_{ij}$, $j = 1,..., N_i$, $\varsigma_{ij} \sim U[0,5]$. Stratify the areas into 3 strata; stratum U_1 consists of areas $1 \le i \le 17$, stratum U_2 of areas $17 < i \le 34$ and stratum U_3 of areas $34 < i \le 50$. The rather complicated formula for generating the auxiliary values guarantees that they are the same in each of the strata, except for the random disturbances ς_{ij}.

2- Generate population random area effects $u_i \sim N(0,\sigma_u^2)$, $i=1,...,M$, $\sigma_u^2=100$.

3- Generate *y*-values using the model (5.1) with $\beta=(1,1)'$, $\sigma_e^2=100$.

In order to avoid extreme selection probabilities, the random effects were truncated at $\pm 2.5\sigma_u$, and similarly for the residuals e_{ij} in (5.1).

4- Select 10 areas from each stratum with probabilities $\pi_i = 10z_i / \sum_{j\in U_h} z_j$ by systematic PPS sampling, where $z_i = \text{int}[1000\times\exp(-u_i/8\sigma_u)]$, thus making the area selection informative.

5- Sample n_i units from selected area i by systematic PPS sampling with probabilities $\pi_{j|i} = n_i z_{ij} / \sum_{k=1}^{N_i} z_{ik}$, where $z_{ij} = \exp\{[-(y_{ij}-x_{ij}'\beta)/\sigma_\varepsilon + \delta_{ij}/5]/3\}$, $\delta_{ij} \sim N(0,1)$. Note that the sampling of units is informative and that the selection probabilities satisfy the relationship (5.2). The area sample sizes are fixed in a given stratum, $n_i=5$ if $i\in U_1$, $n_i=25$ if $i\in U_2$ and $n_i=50$ if $i\in U_3$.

6- Repeat Steps 2-5 10,000 times.

For each sample we computed the following 3 predictors of the area means:

A - *'Ordinary'* small area predictors,

$$\hat{\bar{Y}}_i^O = \frac{1}{N_i}[n_i\bar{y}_i + (N_i-n_i)\hat{u}_i + (N_i\bar{X}_i - n_i\bar{x}_i)'\hat{\beta}_{G\hat{L}S}] \text{ if } i\in s,\ \hat{\bar{Y}}_i^O = \bar{X}_i\hat{\beta}_{G\hat{L}S} \text{ for } i\notin s, \quad (9.1)$$

where $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$, $\hat{u}_i = \hat{\gamma}_i(\bar{y}_i - \bar{x}_i'\hat{\beta}_{G\hat{L}S})$, $\hat{\gamma}_i = \hat{\sigma}_u^2/[\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i]$; $(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ were computed by the method of moments (fitting of constants) and β by Generalized Least Squares with the unknown variances replaced by their estimators; see Prasad and Rao (1990) for details. The predictors $\{\hat{\bar{Y}}_i^O\}$ are the EBLUP predictors of $\bar{Y}_i$ for this model under noninformative sampling.

B- *'Design-based'* estimators,

$$\hat{\bar{Y}}_i^D = \bar{y}_{i,w} + (\bar{X}_i - \bar{x}_{i,w})'\hat{\beta}_{pw} \text{ if } i\in s,\ \hat{\bar{Y}}_i^D = \bar{X}_i'\hat{\beta}_{PW} \text{ for } i\notin s, \quad (9.2)$$

$$(\bar{y}_{i,w}, \bar{x}_{i,w}) = \sum_{j\in s_i} w_{ij}(y_{ij}, x_{ij}) / \sum_{j\in s_i} w_{ij},$$

$$\hat{\beta}_{PW} = [\sum_{i\in s,\ j\in s_i} w_i w_{j|i} x_{ij} x_{ij}']^{-1} \sum_{i\in s,\ j\in s_i} w_i w_{j|i} x_{ij} y_{ij}.$$

The predictor $\hat{\bar{Y}}_i^D$ for $i \notin s$ is not really a 'design based' estimator and is similar to the estimator in (9.1), except that $\hat{\beta}_{G\hat{L}S}$ is replaced by the probability weighted estimator $\hat{\beta}_{PW}$. As discussed in the introduction, design based theory is not suited for the prediction of means in nonsampled areas.

C- The new predictors $\hat{\bar{Y}}_i^N$. The predictors are defined by (5.8) for sampled areas and by (5.12) for nonsampled areas. Note that since the population random effects are normal and because of 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 these predictors. The model parameters $\sigma_u^2, \sigma_e^2, \beta$ have been estimated in the same way as for the estimators in A. The coefficients a, b, k_i indexing the relationship between the weights w_{jli} and the outcome and auxiliary variables were estimated by fitting the model (5.2), using the procedures REG and NLIN in SAS.

In addition to the three sets of predictors we computed also the test statistics developed in Section 8 and the variance estimators of the predictors $\hat{\bar{Y}}_i^N$ developed in Section 6, distinguishing between sampled and nonsampled areas. Since the computation of the variances requires generating bootstrap samples, we restricted this part of the simulation study to 300 samples and 300 bootstrap samples for each sample.

Table 1 shows the empirical prediction bias and root mean square error (RMSE) of the three predictors over the 10,000 simulations, separately for sampled and nonsampled areas. Denote by $\bar{Y}_{tr}$ the true mean of area *t* in simulation *r*, *r*=1…10,000, and let $\hat{\bar{Y}}_{tr}$ represent any of the predictors. Define $D_{tr} = 1$ if area *t* is sampled in simulation *r* and $D_{tr} = 0$ otherwise. For a given area *t*, the prediction bias and RMSE when this area is sampled are computed as,

$$Bias_t = \sum_{r=1}^{10,000} D_{tr}(\hat{\bar{Y}}_{tr} - \bar{Y}_{tr}) / \sum_{r=1}^{10,000} D_{tr} \; ; RMSE_t = \sqrt{\sum_{r=1}^{10,000} D_{tr}(\hat{\bar{Y}}_{tr} - \bar{Y}_{tr})^2 / \sum_{r=1}^{10,000} D_{tr}} \qquad (9.3)$$

The prediction bias and RMSE when area *t* is not sampled are obtained by replacing D_{tr} by $(1-D_{tr})$ in (9.3). The results in Table 1 are averages over the areas contained in the same stratum (having the same sample size). Table 1 shows also the means of the variance estimators developed in Section 6.

Table 1 about here

The conclusions from Table 1 are clear-cut:

1- Ignoring the informative sampling scheme induces large prediction bias for both sampled and nonsampled areas. The large biases induce large RMSEs.

2- The design based estimators are approximately unbiased in sampled areas when the sample sizes within the areas are sufficiently large ($n_i = 25$ in our study), but are biased when estimating the means of nonsampled areas. Recall that no design unbiased predictor for a given nonsampled area exists in general.

3- The new predictors $\hat{\bar{Y}}_i^N$ are literally unbiased for both sampled and nonsampled areas.

4- The RMSEs of all the predictors for sampled areas decrease as the sample sizes within the areas increase.

5- The RMSEs of the predictor $\hat{\bar{Y}}_i^N$ in nonsampled areas are lower than the RMSEs of the other two predictors but they seem high, particularly when compared to the RMSEs obtained for the sampled areas. Note, however, that for nonsampled areas the standard deviation of the random effect is $Std_c(u_i) \cong 9.75$, which is only slightly smaller than the RMSEs of $\hat{\bar{Y}}_i^N$.

6- The proposed RMSE estimates are basically unbiased for both sampled and nonsampled areas.

The magnitude of the bias and the precision of the RMSE estimators can be further assessed by the performance of confidence intervals for the area means derived from them. Table 2 shows the coverage rates of the conventional confidence intervals $\hat{\bar{Y}}_i^N \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{\bar{Y}}_i^N)}$ for $1-\alpha = 0.90, 0.95, 0.99$. The match with the nominal levels is almost perfect.

Table 2 about here

We emphasized in Section 8 the need for testing the informativeness of the sample selection. Notwithstanding, statistical tests have their limitations, and it is important to assess the performance of the new predictors when the sample selection is in fact noninformative. To this end, we sampled the areas with probabilities proportional to a size variable $z_i \sim U[1,2]$ that is independent of the random area effects, and sampled the units within selected area i with probabilities proportional to $z_{ij} = \exp[(x_{ij}/10) + (\varphi_{ij}/5) - 1]$, where $\varphi_{ij} \sim N(0,1)$, independently of y_{ij}. Table 3 shows the bias, RMSE and mean of the RMSE estimators as obtained in this case. Table 4 shows the corresponding coverage rates of confidence intervals for the true means, similar to Table 2.

Table 3 and 4 about here

For noninformative sampling of areas and within the areas, the ordinary small area predictors $\hat{\bar{Y}}_i^O$ are in common use, but the results in Table 3 show that the new predictor, although being much more entangled, performs equally well both in terms of bias and RMSE. The RMSE estimators again perform well, with a positive bias of up to 4% in sampled areas and a negative bias of up to 3% in nonsampled areas. The match between the empirical coverage rates and the nominal levels in Table 4 is again almost perfect.

Table 5 about here

Finally, Table 5 shows the distributions of the test statistics t^A (Eq. 8.1) and $f_{\max}^w$ (Eq. 8.2) designed for testing the informativeness of the sampling of areas and within the areas, for the case where the sampling at both levels is noninformative. The empirical percentiles for both tests are almost identical to the nominal percentiles, despite the fact that the relationships between w_i and u_i, and between $w_{j|i}$ and (y_{ij}, x_{ij}) are nonlinear. Moreover, when applying the tests to the samples obtained by informative sampling, the null hypothesis of noninformative selection of areas was always rejected with p-value lower than 0.01, and the null hypothesis of noninformative sampling within the selected areas was always rejected with p-value lower than 0.025.

10. ESTIMATION OF MEAN BODY MASS INDEX IN USA COUNTIES

10.1 *The sample data*

In this section we apply the methodology developed in the previous sections for estimating the mean body mass index (BMI) for counties in the USA. The BMI is defined as the ratio between the weight, measured in kilograms, and the square of the height, measured in meters. An index higher than 27.8 for men and higher than 27.3 for women is considered as overweight, which is known to be a major health risk factor. Estimating the mean BMI at the national and subnational level is therefore of prime importance for health authorities dealing with this problem. (A new national campaign to fight obesity has just been launched in the UK.) The data used for this study were collected as part of the third national health and nutrition examination survey (NHANES III), and it was provided to us by the national Center for Health Statistics (NCHS). The survey was conducted in two phases during the years 1988-1991 and 1991-1994, and it represents the total civilian noninstitutional population in the US.

NHANES III is a stratified four-stage clustered survey that collects health, dietary and background information through questionnaires and physical examinations. The primary sampling units (PSU) are in most cases individual counties. There are 81 PSUs in the sample, selected with probability proportional to a measure of size without replacement. The size measure was constructed in such a way that the survey oversampled PSUs with large population sizes of Mexican-Americans and Blacks. The second stage of the sample selection consisted of sampling of area segments, which were then stratified based on the percent of Mexican-Americans. Next, households were sampled within the strata, with higher rates for strata with high minority concentrations. In the last stage a sample of persons was selected from classes of households defined by age, sex and race and here again, the classes were sampled at different rates. For more details of the NHANES III sample design and the computation of the sampling weights, see http://www.cdc.gov/nchs/about/major/nhanes/nh3data.htm. The data set used for this study refers to the 81 sampled counties. There are 3138 counties in total in the US. The numbers of sampled persons within the sampled counties are large relative to a typical small area estimation problem, with almost

all the sample sizes exceeding 80. The total sample size is 16,521; 8767 women and 7754 men. Thus, the major small area estimation problem with this survey is that only a small fraction of the counties that define the areas is represented in the sample.

10.2 *Analysis*

In a previous article, Malec *et al*. (1999) used NHANES III data for estimating overweight prevalence for states in the US by fitting logistic models with fixed age/race/gender effects and random race/gender effects. In order to account for sampling effects within the selected counties, the authors estimated the sampling probabilities by utilizing the sampling weights, and then substituted the estimates in the likelihood. The state prevalence estimates were obtained by applying the Bayesian approach with the aid of MCMC simulations.

In our application we fit the model (5.1), separately for men and women, with county random effects and seven covariates. A constant, 3 dummy race variables and 3 age variables. The race variables are: $x_1 = 1$ if non Hispanic white, $x_2 = 1$ if non Hispanic black and $x_3 = 1$ if Hispanic. The age variables are: $x_4 = age \times \mathrm{I}_{20 \le age < 50}$, $x_5 = age \times \mathrm{I}_{50 \le age < 75}$, $x_6 = age \times \mathrm{I}_{75 \le age}$. The age variables are used as proxy for a quadratic relationship between the BMI and age. We could not include age^2 in the model because the true county means of this variable are unknown. There are a few other covariates with sample measurements that affect the BMI but could not be included in the model for the same reason. One of these variables is education, measured by the number of years at school, which was found to have a negative effect on the BMI level of women. The data files that we could use only contain information on the county numbers of adults with college and higher education, but this information is unknown at the individual level.

Table 6 shows the estimated regression coefficients, their standard errors and the estimates of the variance of the random effects and the residual variance. All the coefficients except for the coefficient of 'White non Hispanic' in the women's model are significant, and interesting enough, the variances in the model for women are much larger than in the model for men. We tested the assumption that the residual variance is constant across the counties by first fitting the model

for each of the sampled counties separately (and hence assuming fixed county effects) and then testing the homogeneity of the estimated residuals. After dropping 7 outlying counties, the hypothesis of homogeneity is accepted using Bartlett's test, with p-values of 0.99 for women and 0.13 for men.

Table 6 about here

Next we applied the tests for sample ignorability proposed in Section 8. Unlike Malec *et al*. (1999), we obtain that for both men and women the sampling within the counties is noninformative (given the covariates included in the model), and that the sampling of counties is informative for women, but not for men. The p-values when testing the sampling ignorability within the counties are 0.56 for women and 0.41 for men. The sample ignorability within the counties has been tested also by regressing $\log(w_{j|i})$ against (y_{ij}, x_{ij}) instead of regressing $w_{j|i}$ (see Remark 4 in Section 8.2 and also (5.2)), and by fitting the two regression models in each sampled county separately, confirming in all the cases that for the present model the sample selection within the counties can be ignored. On the other hand, when testing the ignorability of the county selection using (8.1), the p-values are 0.0164 for women and 0.31 for men, suggesting an informative sampling of counties for the women's model but not for the men's model.

As explained in Section 10.1, the sampling probabilities within the counties were determined by the race and age characteristics, and hence it is not surprising that the sampling within the counties was found to be noninformative for the present model that includes race and age as explanatory variables. In this regard, it is not clear how Malec *et al.* (1999) concluded that the sampling within the counties is informative, given that their model likewise accounts for age and race/gender categories. The authors do not elaborate on the reasons for this finding but they show results illustrating different national and state estimates, depending on whether the sampling process is accounted for or not.

The result that the sampling of counties is informative for the women's model is likewise not surprising because the county selection probabilities were determined by the true county race totals and these totals are not included in the model (see below). The model of Malec *et al*. (1999) contains fixed and random race parameters, which is probably why the authors concluded that the selection

of counties is not informative for their model. The fact that the selection of counties was found to be noninformative for the men's model in our application is probably related to the fact that the variance of the county random effects is small, $\hat{\sigma}_u^2 = 0.76$, which makes it harder to detect selection effects.

As mentioned in the introduction, a possible way of controlling sampling effects is by including in the model all the design variables used for the sample selection. In the present application we are in a fortunate (but uncommon) situation where the county design variables; x_{8i} = county total of non Hispanic White, x_{9i} = county total of non Hispanic Black and x_{10i} = county total of Hispanic, are known. Adding these variables (divided by 10^5) to the model yields the following coefficients and standard errors. For women: $\beta_8 = -0.112(0.076)$, $\beta_9 = 0.089(0.200)$, $\beta_{10} = 0.141(0.141)$. For men: $\beta_8 = -0.017(0.043)$, $\beta_9 = -0.064(0.115)$, $\beta_{10} = 0.037(0.079)$. The coefficients and standard errors of the other covariates change only slightly from their values in Table 6 as obtained when fitting the model with only the six covariates. Thus, all three design variables are highly insignificant given the other variables in the model and they are also jointly insignificant with p-values of 0.42 for women and 0.69 for men. With such high p-values, many analysts would tend to drop the design variables from the model and conclude that the sampling of counties is noninformative for the six covariates model, which in view of the previous analysis is not true for the women's model. Furthermore, when re-estimating the random effects using the extended model that includes the three design variables, and applying the informativeness test in (8.1), we find that the sampling of counties is not informative for this model, with p-values 0.17 for women and 0.63 for men. Thus, the selection of counties can only be possibly ignored when including the design variables in the model.

What are the implications of the use of the model with six covariates or the model with 9 covariates (including the 3 design variables) on the prediction of the small area means? In what follows we restrict to the models for women because the selection of counties was found earlier to be noninformative for the men's model even without including the design variables. Starting with the sampled

areas, both models yield very similar predictors when using the predictors defined by (9.1), which are the empirical best linear predictors (EBLUP) under noninformative sampling within the areas ($b = 0$ in (5.8)). For the nonsampled areas, however, they yield somewhat different predictors. Figure 1 shows four different predictors of the means in nonsampled areas. The predictor $\bar{X}_i \hat{\beta}_{G\hat{L}S}$ under the reduced model (6 covariates) as obtained when ignoring the county selection (Eq. (9.1)), the predictor $\bar{X}_i \hat{\beta}_{G\hat{L}S}$ under the extended model with 9 regressors, (the vector $\bar{X}_i$ contains in this case both the proportions and the totals of the three races), the empirical predictor (5.12) under the reduced model (with $b \neq 0$), and the predictor (5.12) under the extended model. The horizontal line at 27.3 marks the threshold defining overweight. For the predictor (5.12) under the reduced model the bias correction, $\sum_{i \in s} (w_i - 1)\hat{u}_i / \sum_{i \in s} (w_i - 1)$ is 0.47 with estimated Jackknife standard deviation of 0.16. For the predictor under the extended model the bias correction is 0.25, with similar estimated standard deviation. Thus, the use of the bias correction for nonsampled areas in the case of the extended model is questionable, in correspondence with the testing result that the selection of counties is noninformative for this model. The use of the Jackknife method for variance estimation assumes that the random effects $\hat{u}_i$ are approximately independent. It is used here only as a rough measure for assessing the stability of the bias correction.

The 4 plots in Figure 1 suggest that ignoring the county selection process and just using the synthetic predictor based on the 6 regressors' model under-predicts the true county means. This becomes evident by comparing the synthetic predictors under this model with the synthetic predictors obtained under the extended model. The latter predictors are lower than the predictors obtained under the 6 covariates model with the bias correction, but interesting enough, once the bias correction is added also to the predictors under the extended model, both sets of predictors behave very similarly. However, as discussed above, the use of a bias correction for the extended model is questionable.

The magnitudes of the bias corrections seem very small, but they are not negligible. To see this, we computed the percentages of nonsampled areas for which the predicted means are higher than the threshold of 27.3, as obtained by use of the four predictors. The use of the two synthetic predictors yields a percentage of 2.84% for the six covariates model and 5.56% for the extended model. Adding the bias correction of 0.47 to the first synthetic predictor changes the percentage to 9.2%, whereas adding the bias correction of 0.25 to the second synthetic predictor changes the percentage to 10.3%. Thus, if areas with means that exceed the threshold are to be given extra attention, the use of the bias correction can be very important.

11. CONCLUDING REMARKS

This article presents a first attempt of predicting small area means under informative sampling of areas and within the areas. The proposed procedure assumes knowledge of the models holding for the sample data and for the sampling weights within the selected areas, but otherwise is 'model free'. Both models can be identified and estimated from the sample data. In the present application we consider the familiar nested error regression model but as outlined in Section 7, the procedure can be applied to other models with continuous or discrete outcomes using similar steps. Note, in particular, that for the familiar Fay and Herriot (1979) model the input data consists of unbiased design based estimators for the area means or proportions, so that in this case one only needs to account for the informativeness of the area selection.

Much of the research in small area estimation concerns the use of Bayesian methods that allow considering heavy structured models and accounting for all sources of variation when assessing the prediction errors. In this article we restrict to the frequentist approach but it would seem that the proposed procedure can be adapted and used in a Bayesian set up, except that it will require modelling the relationship between the area selection probabilities and the true area means, which as discussed in Section 5 is more complicated but not necessary under the present procedure. See Pfeffermann *et al*. (2006) for an example of modelling this relationship. Developing a Bayesian solution that does not require this extra step is an intriguing problem.

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), "An 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.

Hall, P. and Maiti, T. (2006), "On parametric bootstrap methods for small area predictions," *Journal of the Royal Statistical Society* 68, Series B, 221-238.

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.

Malec, D., Davis, W. W., and Cao, X. (1999). "Model-based small area estimates of overweight prevalence using sample selection adjustment," *Statistics in Medicine* 18, 3189-3200.

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., 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.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P. L. (2006), "Multilevel modeling under informative sampling," *Biometrika*, 93, 943-959.

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

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.

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. (2004), "Prediction of finite population totals based on the sample distribution," *Survey Methodology,* 30, 79-92.

Table 1. Bias, Root Mean Square Error (RMSE) and mean of RMSE estimators (RMSE-E). Informative sampling of areas and within areas

		Sampled Areas			Nonsampled Areas		
	Sample size	Ordinary $\hat{\bar{Y}}_i^O$	Design $\hat{\bar{Y}}_i^D$	New $\hat{\bar{Y}}_i^N$	Ordinary $\hat{\bar{Y}}_i^O$	Design $\hat{\bar{Y}}_i^D$	New $\hat{\bar{Y}}_i^N$
Bias	$n_i = 5$	-3.25	-0.71	-0.02	-6.36	-2.00	-0.32
	$n_i = 25$	-3.27	-0.14	-0.09	-6.10	-1.73	-0.06
	$n_i = 50$	-3.27	-0.07	-0.15	-6.10	-1.73	-0.06
RMSE	$n_i = 5$	5.26	4.88	4.14	11.77	10.04	9.85
	$n_i = 25$	3.80	2.19	1.95	11.70	10.08	9.93
	$n_i = 50$	3.54	1.54	1.39	11.71	10.11	9.96
RMSE-E	$n_i = 5$	---	---	4.28	---	---	9.90
	$n_i = 25$	---	---	2.02	---	---	9.91
	$n_i = 50$	---	---	1.46	---	---	9.91

Table 2. Coverage rates of confidence intervals for true area means. Informative sampling of areas and within areas

		Sampled Areas			Nonsampled Areas		
	Nominal levels	0.90	0.95	0.99	0.90	0.95	0.99
Sample size	$n_i = 5$	0.90	0.94	0.98	0.91	0.95	0.99
	$n_i = 25$	0.89	0.94	0.98	0.91	0.95	0.99
	$n_i = 50$	0.89	0.94	0.98	0.92	0.96	0.99

Table 3. Bias, Root Mean Square Error (RMSE) and mean of RMSE estimators (RMSE-E). Noninformative sampling of areas and within areas

		Sampled Areas			Nonsampled Areas		
	Sample size	Ordinary $\hat{\bar{Y}}_i^O$	Design $\hat{\bar{Y}}_i^D$	New $\hat{\bar{Y}}_i^N$	Ordinary $\hat{\bar{Y}}_i^O$	Design $\hat{\bar{Y}}_i^D$	New $\hat{\bar{Y}}_i^N$
Bias	$n_i = 5$	-0.03	-0.02	-0.02	0.17	0.18	0.19
	$n_i = 25$	0.00	0.00	0.01	-0.02	-0.04	-0.01
	$n_i = 50$	-0.02	-0.01	-0.01	0.01	0.01	0.02
RMSE	$n_i = 5$	4.12	4.45	4.12	10.23	10.69	10.17
	$n_i = 25$	1.96	1.99	1.96	10.16	10.10	10.12
	$n_i = 50$	1.34	1.39	1.38	10.27	10.21	10.24
RMSE-E	$n_i = 5$	---	---	4.30	---	---	9.89
	$n_i = 25$	---	---	2.03	---	---	9.96
	$n_i = 50$	---	---	1.46	---	---	9.93

Table 4. Coverage rates of confidence intervals for true area means. Noninformative sampling of areas and within areas

		Sampled Areas			Nonsampled Areas		
	Nominal levels	0.90	0.95	0.99	0.90	0.95	0.99
Sample size	$n_i = 5$	0.89	0.94	0.99	0.91	0.95	0.99
	$n_i = 25$	0.89	0.94	0.99	0.91	0.96	0.99
	$n_i = 50$	0.89	0.95	0.99	0.91	0.96	0.99

Table 5. Distribution of test statistics for testing the sampling informativeness under noninformative sampling of areas and within the areas

Percentiles	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
Sampling of areas	0.013	0.029	0.053	0.107	0.896	0.952	0.975	0.988
Sampling within areas	0.009	0.025	0.049	0.093	0.903	0.948	0.976	0.988

Table 6. Regression coefficients, standard errors (in parentheses) and variances for BMI models fitted to data from NHANES III

Coeff.	Intercept	White Non Hispanic	Black Non Hispanic	Hispanic	Age<50	50≤Age<75	Age≥75
Men	22.960 (0.414)	0.739 (0.314)	0.740 (0.316)	1.161 (0.322)	0.083 (0.008)	0.056 (0.005)	0.020 (0.004)
Women	21.852 (0.526)	-0.670 (0.374)	2.355 (0.375)	1.602 (0.394)	0.133 (0.010)	0.095 (0.006)	0.049 (0.005)

Men: $\sigma_u^2 = 0.760$, $\sigma_e^2 = 23.040$; *Women*: $\sigma_u^2 = 2.830$, $\sigma_e^2 = 39.560$

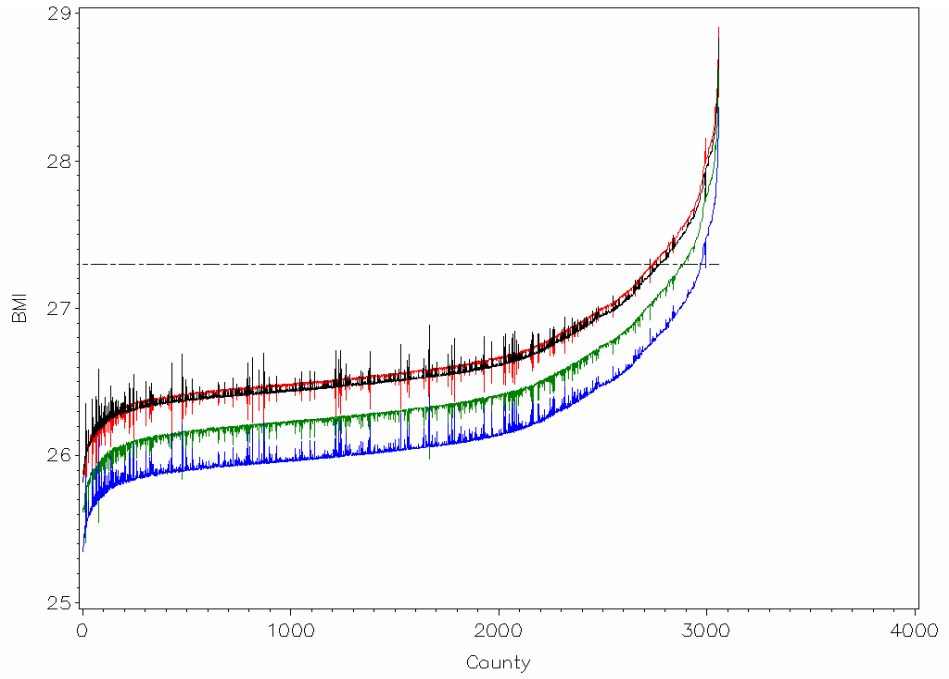

Figure 1. Prediction of mean body mass index of women in nonsampled counties of NHANES III. Values above the horizontal line at 27.3 define 'overweight'.

The blue and green lines show the synthetic predictors under the six covariates model and the 9 covariates model respectively. The dark and red lines show the corresponding predictors with bias corrections. The counties are ordered by the average values of the 4 predictors.