SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING AND NOT MISSING AT RANDOM NONRESPONSE December 2018

Michael Sverchkov, Bureau of Labor Statistics, Washington DC, USA
Danny Pfeffermann, Government Statistician of Israel; *Professor*: *Hebrew University of Jerusalem, Israel* & *University of Southampton, UK*. (**)

ABSTRACT

Pfeffermann and Sverchkov (P-S 2007) considered Small Area Estimation (SAE) for the case where the selection of the sampled areas is informative in the sense that the area sampling probabilities are related to the true (unknown) area means, and the sampling of units within the selected areas is likewise informative with probabilities that are related to the values of the study variable; in both cases after conditioning on the model covariates. In this paper we extend this approach to the practical situation of incomplete response at the unit level, and where the response is not missing at random (NMAR). The proposed extension consists of first identifying the model holding for the observed responses and using the model for estimating the response probabilities, and then applying the approach of P-S to the observed data with the unit sampling probabilities replaced by the products of the sampling probabilities and the estimated response probabilities. A bootstrap procedure for estimating the MSE of the proposed predictors is developed. We illustrate our approach by a simulation study and by application to a real data set. The simulations also illustrate the consequences of not accounting for informative sampling and nonresponse.

Key words: missing information principle (MIP), population distribution, respondents model, sample distribution.

Acknowledgment: We thank Dan Benhur from the Central Bureau of Statistics in Israel for many valuable comments on the theory and computations of this article.

(**) *The opinions expressed in this paper are of the authors and do not necessarily represent the policies of the U.S. Bureau of Labor Statistics and the Israel Central Bureau of Statistics.*

1. INTRODUCTION

Over the last 20 years, many articles have been published on how to account for informative sampling when estimating population parameters from informative probability samples. See Pfeffermann and Sverchkov (2009), Pfeffermann (2011) and Kim and Skinner (2013) for reviews and discussion. By informative sampling we mean that the sampling probabilities are related to the outcome variable of interest, even after conditioning on model covariates, such that the conditional distribution of the study variable given the covariates differs from the corresponding distribution in the population from which the sample is taken. As illustrated in the literature and also in the empirical study of the present article, not accounting for informative sampling and nonresponse, can result in large bias and root MSE (RMSE), and hence in misleading inference.

In the last decade, several approaches have been proposed to deal with informative sampling in the context of small area estimation (SAE). See Pfeffermann (2013) for a review of methods. In particular, Pfeffermann and Sverchkov (2007) considered the case where the selection of the sampled areas is informative in the sense that the area sampling probabilities are related to the true (unknown) area means, and the sampling of units within the selected areas is likewise informative, with probabilities that are related to the values of the study variable, in both cases after conditioning on the model covariates. Verret et al. (2015) proposed an alternative method to account for informative sampling within the sampled areas. We consider the approach of Pfeffermann and Sverchkov (2007) later in this article, using an important result of Verret et al. (2015).

A related, but definitely more complicated problem when analyzing survey data is not missing at random (NMAR) nonresponse. Here the problem is that no information is obtained from some of the sampled units, with the propensity to respond possibly depending on the study variable of interest, even after conditioning on known covariates. As is well known, response rates have dropped very drastically over the years, sometimes being even lower than 50%. The obvious reason why this is a much more complicated problem is that unlike the sampling probabilities in informative sampling, the response probabilities are generally unknown and cannot readily be estimated from the observed data since the missing data are unobserved, requiring one to assume some structure for

these probabilities. Because NMAR nonresponse is such a complicated problem, analysts often assume either explicitly or implicitly the existence of covariates known for all the sample elements, which explain the response probabilities in the sense that after conditioning on these covariates, the probability to respond no longer depends on the study variable, commonly known as missing at random (MAR). It is far beyond the scope of this article to review all the rich literature devoted to this theme. See Pfeffermann and Sikov (2011) and Riddles et al. (2016) for reviews and references.

The primary objective of the present article is to propose a method of handling NMAR nonresponse in the framework of SAE. Notice that in official statistics, the sample used for SAE is basically the same sample used to obtain direct national or subnational estimates (areas or domains with large samples for which the estimates are based on only the data observed for them). Consequently, the reasons for nonresponse are the same in both cases, although the problems resulting from the nonresponse can be more severe in SAE because of the small sample sizes within at least some of the areas, even under full response. To the best of our knowledge, no article has been published considering this very important problem of NMAR nonresponse. To this end, we extend the approach of Pfeffermann and Sverchkov (2007). The proposed extension consists of identifying the outcome model holding for the observed responses and using this model for estimating the response probabilities by application of the Missing Information Principle (MIP). For this, we define the likelihood holding for the sample under complete response, we integrate out the unobserved outcomes from this likelihood over the outcome distribution holding for the nonrespondents, and then solve the resulting likelihood equations. Having estimated the response probabilities, we apply the approach of Pfeffermann and Sverchkov (2007) to the observed data for the respondents, with the unit sampling probabilities replaced by the products of the sampling probabilities and the estimated response probabilities.

The paper is organized as follows: In Section 2 we introduce the basic notation and define the models holding for the responding and the non-responding sample units. In Section 3 we outline the basic steps of our proposed approach for estimating the response probabilities. Section 4 considers two alternative ways of estimating the small area means

once the response probabilities have been estimated, namely, the use of direct estimates and the use of empirical model-based estimators. Prediction MSE estimation is considered in Section 5, followed by a simulation study in Section 6, aimed to illustrate the performance of our proposed predictors in comparison to predictors that ignore the informative sampling process or the NMAR nonresponse mechanism. The proposed procedure is applied to a real data set from Israel in Section 7. We conclude with a brief summary in Section 8.

2. NOTATION AND MODELS

Let $\{y_{ij}, \mathrm{x}_{ij}; i = 1,..., M,\ j = 1,..., N_i\}$ represent the data in a finite population of N units belonging to M areas with N_i units in area i, $\sum_{i=1}^{M} N_i = N$, where y_{ij} is the value of the study variable for unit j in area i and $\mathrm{x}'_{ij} = (x_{ij,1},..., x_{ij,K})$ is a vector of corresponding K covariates. We assume that the covariates are known for every unit in the population. Suppose that the outcome values follow the generic two-level population model:

$$\begin{aligned} & y_{ij} \mid \mathrm{x}_{ij}, u_i^U \sim f(y_{ij} \mid \mathrm{x}_{ij}, u_i^U),\ i = 1,..., M,\ j = 1,..., N_i \\ & u_i^U \sim f(u_i^U);\ E(u_i^U) = 0,\ V(u_i^U) = \sigma^2_{u^U}, \end{aligned} \tag{2.1}$$

where u_u^U is the i[th] area level random effect under this model. The target is to estimate the area means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}, i = 1,..., M$, based on a sample obtained by the following two-stage sampling scheme: select a sample s of m out of the M population areas with inclusion probabilities $\pi_i = \Pr(i \in s)$; select a sample s_i of $n_i > 0$ units from selected area i with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$. Denote by I_i, I_{ij} the sample indicators- $I_i = 1$ if area i is selected in the first stage and 0 otherwise, $I_{ij} = 1$ if unit j of selected area i is sampled in the second stage and $I_{ij} = 0$ otherwise. Let $w_i = 1/\pi_i$, $w_{j|i} = 1/\pi_{j|i}$ denote the first- and second-stage sampling weights.

In practice, not every unit in the sample responds. Define the response indicator $R_{ij} = 1$ if unit $j \in s_i$ responds and $R_{ij} = 0$ otherwise. The sample of respondents is thus

$R = \{(i, j) : I_i = 1, I_{ij} = 1, R_{ij} = 1\}$ and the sample of nonrespondents among the sampled units is $R^c = \{(i,k) : I_i = 1, I_{ik} = 1, R_{ik} = 0\}$. We assume $\sum_{j=1}^{n_i} R_{ij} > 0$ for all the sampled areas. The sample of respondents can thus be viewed as the result of a two-stage sampling process where in the first stage the sample is selected from the population with known inclusion probabilities, and in the second stage the sample is "self-selected" with unknown response probabilities (Särndal and Swensson, 1987).

Define, $u_i = u_i^U - E(u_i^U \mid i \in s)$. Then, under the population model (2.1), the observed data follow the two-level 'respondents' model:

$$y_{ij} \mid x_{ij}, u_i \sim f_R(y_{ij} \mid x_{ij}, u_i) = f(y_{ij} \mid x_{ij}, u_i, (i, j) \in R);\ u_i \sim f(u_i \mid i \in s), E(u_i \mid i \in s) = 0\,. \quad (2.2)$$

The model in (2.2) is again general and all that we state at this stage is that under informative sampling and/or NMAR nonresponse, the population and the respondents models differ, $f_R(y_{ij} \mid x_{ij}, u_i) \neq f(y_{ij} \mid x_{ij}, u_i^U)$.

Remark 1. The respondents' model refers to the observed data and hence can be estimated and tested by standard small area estimation (SAE) methods. See Pfeffermann (2013) and Rao and Molina (2015) for estimation and testing procedures in SAE, with references.

Let $p_r(y_{ij}, x_{ij}) = \Pr[R_{ij} = 1 \mid y_{ij}, x_{ij}, i \in s, j \in s_i]$. If the probabilities $p_r(y_{ij}, x_{ij})$ were known, the sample of respondents could be considered as a two-stage sample from the finite population with known selection probabilities π_i and $\tilde{\pi}_{j|i} = \pi_{j|i} p_r(y_{ij}, x_{ij})$. Also, if known, the response probabilities could be used for imputation of the missing data within the selected areas, by application of the relationship between the sample and sample-complement distributions, (Sverchkov and Pfeffermann, 2004),

$$f(y_{ij} \mid x_{ij}, u_i, (i, j) \in R^c) = \frac{[p_r^{-1}(y_{ij}, x_{ij}) - 1] f(y_{ij} \mid x_{ij}, u_i, (i, j) \in R)}{E\{[p_r^{-1}(y_{ij}, x_{ij}) - 1] \mid x_{ij}, u_i, (i, j) \in R\}}\,. \quad (2.3)$$

Notice that under informative NMAR nonresponse, the nonrespondents' distribution differs from the respondents' distribution $f(y_{ij} \mid x_{ij}, u_i, (i, j) \in R)$. As stated in Remark 1, the latter distribution refers to the observed data and therefore can be fitted by classical

SAE methods, allowing in turn estimating the nonrespondents' distribution via (2.3). In the following section we show how we can estimate the response probabilities.

3. ESTIMATION OF RESPONSE PROBABILITIES

In what follows we assume a parametric model for the response probabilities, which depends on an unknown vector parameter γ; $p_r(y_{ij}, x_{ij}) = p_r(y_{ij}, x_{ij}; \gamma)$ $= \Pr[R_{ij} = 1 | y_{ij}, x_{ij}, i \in s, j \in s_i; \gamma]$.

Assumption 1. $p_r(y_{ij}, x_{ij}; \gamma)$ is differentiable with respect to γ, and the response probabilities (but not necessarily the second stage sample sampling probabilities) are independent between the units; $p_r(y_{ij}, y_{ik}, x_{ij}, x_{ik}; \gamma) = p_r(y_{ij}, x_{ij}; \gamma) p_r(y_{ik}, x_{ik}; \gamma)$.

Assumption 2. $f(y_{ik} | O, u_i, (i,k) \in R^c) = f(y_{ik} | x_{ik}, u_i, (i,k) \in R^c)$, where O represents all the observed data; $O = \{y_{ij}, \pi_{j|i}, \pi_i, n_i, (i, j) \in R;\ x_{hl}, h = 1,..., M,\ l = 1,..., N_i\}$. The assumption states that the unobserved outcomes in a sampled area are independent of the observed outcomes, given the area random effect and the covariates. Pfeffermann and Sverchkov (2007) define two general mild conditions under which the assumption holds (adapted to the present context of NMAR nonresponse):

(C1) $f[y_{il}, y_{ij} | u_i, x_{il}, x_{ij}, (i,l) \notin R, R_{ij} = 1] = f[y_{il} | u_i, x_{il}, (i,l) \notin R] f(y_{ij} | u_i, x_{ij}, R_{ij} = 1)$,

(C2) $f(\pi_{j|i} | u_i, y_{il}, y_{ij}, x_{ij}, x_{il}, (i,l) \notin R, R_{ij} = 1) = f(\pi_{j|i} | u_i, y_{ij}, x_{ij}, R_{ij} = 1)$

The first condition is very mild since the outcomes in a given area are independent given the random effect, and the area selection probability is related to the area mean and not to individual deviations from the mean, such that by conditioning on the random effect the independence of the outcomes is preserved. The second condition also seems mild for the common situation in small area estimation of large true area sizes but small samples.

Under these assumptions, if the missing outcome values were actually observed, γ could be estimated by solving the likelihood equations:

$$\sum_{(i,j) \in R} \frac{\partial \log p_r(y_{ij}, x_{ij}; \gamma)}{\partial \gamma} + \sum_{(i,k) \in R^c} \frac{\partial \log[1 - p_r(y_{ik}, x_{ik}; \gamma)]}{\partial \gamma} = 0. \qquad (3.1)$$

In practice, the missing data are unobserved and hence the likelihood equations (3.1) are not operational. However, one may apply in this case the missing information principle:

Missing Information Principle (Cepillini et al. 1955, Orchard and Woodbury, 1972)**:** since no observations are available for $(i,k) \in R^c$, solve instead,

$$E_U \left\{ \left[\sum_{(i,j)\in R} \frac{\partial \log p_r(y_{ij}, x_{ij};\gamma)}{\partial \gamma} + \sum_{(i,\mathrm{k})\in R^c} \frac{\partial \log[1-p_r(y_{ik}, x_{ik};\gamma)]}{\partial \gamma} \right] \middle| O \right\}$$

$$= \sum_{(i,j)\in R} \frac{\partial \log p_r(y_{ij}, x_{ij};\gamma)}{\partial \gamma} + \sum_{(i,k)\in R^c} E_{nre} \left\{ \frac{\partial \log[1-p_r(y_{ik}, x_{ik};\gamma)]}{\partial \gamma} \middle| O, (i,k) \in R^c \right\}$$

$$= \sum_{(i,j)\in R} \frac{\partial \log p_r(y_{ij}, x_{ij};\gamma)}{\partial \gamma}$$

$$+ \sum_{(i,k)\in R^c} E_s \left\{ E_{nre} \left[\frac{\partial \log[1-p_r(y_{ik}, x_{ik};\gamma)]}{\partial \gamma} \middle| O, \mathrm{u}_i, (i,k) \in R^c \right] \middle| O, (i,k) \in R^c \right\}$$

= by (2.3) and Assumption 2,

$$\sum_{(i,j)\in R} \frac{\partial \log p_r(y_{ij}, x_{ij};\gamma)}{\partial \gamma}$$

$$+ \sum_{(i,k)\in R^c} E_s \left(\frac{E_{re} \left\{ [p_r^{-1}(y_{ik}, x_{ik};\gamma) - 1] \frac{\partial \log[1-p_r(y_{ik}, x_{ik};\gamma)]}{\partial \gamma} \middle| \, x_{ik}, u_i, (i,k) \in R \right\}}{E_{re}\{[p_r^{-1}(y_{ik}, x_{ik};\gamma) - 1] \,|\, x_{ik}, u_i, (i,k) \in R\}} \middle| O \right) = 0 \,. \quad (3.2)$$

In (3.2) $E_U, \mathrm{E}_s, \mathrm{E}_{re}, E_{nre}$ define respectively expectations with respect to the population distribution, the sample distribution, the respondents' distribution and the non-respondents' distribution. Notice that the internal expectations in the last expression are with respect to the model holding for the observed data for the respondents.

The rationale of the MIP is simple. Ideally, we would want to use the score function (3.1), but since the outcomes are unknown for the nonresponding units, we replace the second expression, $\sum_{(i,\mathrm{k})\in R^c} \frac{\partial \log[1-p_r(y_{ik}, x_{ik};\gamma)]}{\partial \gamma}$ by its "best predictor", as defined by its

expectation given the observed data; $\sum_{(i,k)\in R^c} E_{nre}\left\{\frac{\partial \log[1-p_r(y_{ik},x_{ik};\gamma)]}{\partial\gamma}\middle| O,(i,k)\in R^c\right\}$.

Orchard and Woodbury (1972) formalize this step more generally as follows: denote by $f(O,M;\theta)$ the joint distribution of the observed and missing data, indexed by the vector parameter θ. If the missing data were actually observed, one could estimate θ by the score function obtained from f. (The score (3.1) in the present case.) But since M is unobserved, factorize $f(O,M;\theta)=f_1(O;\theta)f_2(M\,|\,O;\theta)$ and estimate θ from the marginal distribution f_1 of the observed data, and the expectation $E_2(M\,|\,O;\theta)$.

Returning to the MIP equations in (3.2), the vector parameter γ is estimated by replacing u_i by $\hat{u}_i$ and dropping the external expectation. In our empirical study we solved the resulting equations by minimizing the log-likelihood leading to them, i.e., minimizing,

$$\sum_{(i,j)\in R}\log p_r(y_{ij},x_{ij};\gamma)$$
$$+\sum_{(i,k)\in R^c} E_s\left(\frac{E_{re}\{[p_r^{-1}(y_{ik},x_{ik};\gamma^*)-1]\log[1-p_r(y_{ik},x_{ik};\gamma)]\,|\,x_{ik},u_i,(i,k)\in R\}}{E_{re}\{[p_r^{-1}(y_{ik},x_{ik};\gamma^*)-1]\,|\,x_{ik},u_i,(i,k)\in R\}}\middle| O\right). \qquad (3.3)$$

We distinguish between γ^* and γ because by (3.2), the derivatives should only be taken with respect to γ. The minimization was thus carried out iteratively by minimizing on the (q+1) iteration the function,

$$\sum_{(i,j)\in R}\log p_r(y_{ij},x_{ij};\gamma^{(q+1)})$$
$$+\sum_{(i,k)\in R^c} E_s\left(\frac{E_{re}\{[p_r^{-1}(y_{ik},x_{ik};\gamma^{(q)})-1]\log[1-p_r(y_{ik},x_{ik};\gamma^{(q+1)})]\,|\,x_{ik},u_i,(i,k)\in R\}}{E_{re}\{[p_r^{-1}(y_{ik},x_{ik};\gamma^{(q)})-1]\,|\,x_{ik},u_i,(i,k)\in R\}}\middle| O\right) \qquad (3.4)$$

with respect to $\gamma^{(q+1)}$. The use of this procedure worked well in our empirical study, but other numerical procedures can possibly be considered for solving the estimating equations resulting from (3.2).

Remark 2. When the response probabilities $p_r(y_{ij},x_{ij};\gamma)$ depend on only x_{ij} (and γ), they are referred to as *propensity scores*, and the missing data are missing at random. This kind of response mechanism may hold in establishment survey settings, for example,

when the response propensity is related to the unit size. The estimating equations in (3.2) reduce in this case to the common log-likelihood equations,

$$\sum_{(i,j)\in R} \frac{\partial \log p_r(x_{ij};\gamma)}{\partial \gamma} + \sum_{(i,k)\in R^c} \frac{\partial \log[1-p_r(x_{ik};\gamma)]}{\partial \gamma} = 0, \tag{3.5}$$

where $p_r(x_{ij};\gamma) = \Pr(\mathrm{R}_{ij} = 1 | x_{ij};\gamma)$.

Remark 3: A fundamental question regarding the solution of the MIP equations (3.2) is the existence of a unique solution, or more generally, the identifiability of the response model. In a recent article, Riddles et al. (2016) proposed a similar approach to deal with NMAR nonresponse in the general context of sample surveys and established the following fundamental condition for the response model identifiability: the covariates x can be decomposed as $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2)$ with $dim(\mathrm{x}_2) \geq 1$, such that $\Pr(R_{ij} = 1 | y_{ij}, x_{ij}) = \Pr(R_{ij} = 1 | y_{ij}, x_{1ij})$. In other words, the covariates in x_2 that appear in the outcome model do not affect the response probabilities, given the outcome and the other covariates. Although not explaining the response, the variables in x_2 explain the variability of the outcome values and hence they provide valuable information on the missing values, and are therefore essential for estimating the parameters underlying the response mechanism.

Variable(s) of this property may or may not exist in a general set up, but interesting enough, SAE models actually contain such a variable, namely, the random effects. The random effects play a fundamental role in SAE models so the outcome clearly depends on them, but it is reasonable to assume that the response probabilities do not depend on the random effect given the outcome value. In practice, the random effects are unobservable but we estimate them and then solve the equations (3.2) by conditioning on the estimated effects. So, it is actually the estimated random effects that play the role of the covariates x_2. (Other covariates that are predictive of the outcome but not of the response might exist as well). Clearly, the larger the absolute values of the random effects, the more they affect the values of the outcome values and hence also the values of the response probabilities. In the simulation study of Section 6 we study the effect of the magnitude of the variance of the random effects on the prediction of the area means.

Example: *Mixed logistic model for outcome variable*

Suppose that the model fitted to the observed data of the respondents is the mixed generalized logistic model,

$$p_y(x_{ij},u_i)=\Pr(y_{ij}=1\mid x_{ij},u_i,(i,j)\in R;\beta)=\frac{\exp(\beta_0+\beta_1x_{ij}+u_i)}{1+\exp(\beta_0+\beta_1x_{ij}+u_i)},\ u_i\overset{\text{i.i.d.}}{\sim}N(0,\sigma_u^2). \quad (3.6)$$

Consider a generic response model, $p_r(y_{ij},x_{ij};\gamma)=\Pr[R_{ij}=1\mid y_{ij},x_{ij},i\in s,j\in s_i;\gamma]$.

The components of (3.2) can be written in this case as,

$$E_{re}\left\{[p_r^{-1}(y_{ij},x_{ij};\gamma)-1]\frac{\partial\log[1-p_r(y_{ij},x_{ij};\gamma)]}{\partial\gamma}\middle| x_{ij},u_i,(i,j)\in R\right\}=$$

$$p_y(x_{ij},u_i)[p_r^{-1}(1,x_{ij};\gamma)-1]\frac{\partial\log[1-p_r(1,x_{ij};\gamma)]}{\partial\gamma}+$$

$$[1-p_y(x_{ij},u_i)][p_r^{-1}(0,x_{ij};\gamma)-1]\frac{\partial\log[1-p_r(0,x_{ij};\gamma)]}{\partial\gamma};$$

$$E_{re}\{[p_r^{-1}(y_{ij},x_{ij};\gamma)-1]\mid x_{ij},u_i,(i,j)\in R\}=p_y(x_{ij},u_i)[p_r^{-1}(1,x_{ij};\gamma)-1]+$$

$$[1-p_y(x_{ij},u_i)][p_r^{-1}(0,x_{ij};\gamma)-1].$$

The random effects u_i and the logistic probabilities $p_y(x_{ij},\mathrm{u}_i)$ can be estimated by use of the SAS procedure PROC NLMIX.

Remark 4. A possible criticism of our proposed approach is that it requires specifying a parametric model for the response as a function of the outcome and the covariates, but in general, the model cannot be tested by use of the observed data since the outcomes are missing for the nonrespondents. While this is generally true, we mention that Rivers (2007) and Feder and Pfeffermann (2015) define conditions under which if the true response model is a continuous function of the outcome and the covariates, it can be approximated arbitrarily close by a logistic model with polynomials of the outcome and the covariates, and products of them as the explanatory variables. These results suggest using the logistic model with polynomials and cross products of appropriate orders as the response model. We partly illustrate the robustness of the logistic model as an approximation for the true response probabilities in the simulation study of Section 6. Note again that unlike with the use of the standard propensity scores, which are functions

of only the covariates, the outcome variable is added to the covariates as an additional explanatory variable in the response model, thus accounting for NMAR nonresponse.

4. PREDICTION OF SMALL AREA MEANS

As noted earlier, once the unit level response probabilities have been estimated, the sample of respondents can be considered as a two-stage sample from the finite population with first- and second level estimated probabilities π_i and $\tilde{\pi}_{k|i} = \pi_{k|i} = \pi_{k|i}\, p_r(y_{ik}, x_{ik}; \hat{\gamma})$. By Pfeffermann and Sverchkov (2007), the optimal small area predictor for area i is,

$$\bar{Y}_i^* = E_U(\bar{Y}_i \mid O, I_i). \tag{4.1}$$

(Follows from the identity, $E_U[(\hat{\bar{Y}}_i - \bar{Y}_i)^2 \mid O, I_i) = [\hat{\bar{Y}}_i - E(\bar{Y}_i \mid O, I_i)]^2 + Var_U(\bar{Y}_i \mid O, I_i)$, for any predictor $\hat{\bar{Y}}_i$.) We estimate therefore the area means in *sampled areas* as,

$$\hat{\bar{Y}}_i = \hat{E}_U(\bar{Y}_i \mid O, I_i = 1) = N_i^{-1}[\sum_{j:(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} \hat{E}_P(y_{ik} \mid O, I_i = 1)]$$

$$= N_i^{-1}\left[\sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} E_s\left\{\frac{E_{re}[(\tilde{\pi}_{k|i}^{-1} - 1)y_{ik} \mid x_{ik}, u_i, (i,k)\in R]}{E_{re}[(\tilde{\pi}_{k|i}^{-1} - 1) \mid x_{ik}, u_i, (i,k)\in R]}\middle| O\right\}\right]$$

$$= N_i^{-1}\left[\sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} E_s\left\{\frac{E_{re}\{[\tilde{w}(y_{ik}, x_{ik}) - 1]y_{ik} \mid x_{ik}, u_i, (i,k)\in R\}}{E_{re}\{[\tilde{w}(y_{ik}, x_{ik}) - 1] \mid x_{ik}, u_i, (i,k)\in R\}}\middle| O\right\}\right], \tag{4.2}$$

where $\tilde{w}(y_{ik}, x_{ik}) = E_{re}[\tilde{\pi}_{k|i}^{-1} \mid y_{ik}, x_{ik}, (i,k)\in R]$. (The second row follows from (2.3). We assume $E_{re}[\tilde{\pi}_{k|i}^{-1} \mid y_{ik}, x_{ik}, u_i, (i,k)\in R] = E_{re}[\tilde{\pi}_{k|i}^{-1} \mid y_{ik}, x_{ik}, (i,k)\in R]$). The external expectation in the last row of (4.2) is over the distribution of u_i under the sample model. (No nonresponse of areas). The internal expectations refer to the observed data and therefore can be estimated either by regression or non-parametrically. See Pfeffermann and Sverchkov (2007, 2009), Pfeffermann (2011) and Feder and Pfeffermann (2015) for examples. In Section 6.1 we describe how we estimate the expectations in the empirical study.

Remark 5. The non-responding sampled units in (4.2) are treated the same as non-sampled units. As explained at the beginning of this section, we consider the sample of

respondents as a two-stage sample from the finite population with first- and second level estimated probabilities π_i and $\tilde{\pi}_{k|i} = \pi_{k|i}\hat{p}_r(y_{ik}, x_{ik}; \gamma) = \pi_{k|i} p_r(y_{ik}, x_{ik}; \hat{\gamma})$.

Having estimated the response probabilities, an alternative, almost direct, and in fact simpler predictor of the area mean in a sampled area is the (pseudo) Hajek-Brewer (Hajek, 1971) estimator,

$$\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i}). \qquad (4.3)$$

The prominent feature of this estimator is that it uses the estimated probabilities $\tilde{\pi}_{j|i} = \pi_{j|i} p_r(y_{ij}, x_{ij}; \hat{\gamma})$. As illustrated in the empirical study, this estimator is approximately design-unbiased (it is a Ratio estimator), but with larger sampling variance than the predictor (4.2), particularly in areas with small sample size.

We estimate the area means of the outcomes in *non-sampled areas* as,

$$\hat{\bar{Y}}_i = E_U(\bar{Y}_i \mid O, I_i = 0) = N_i^{-1}[\sum_{k=1}^{N_i} E_U(y_{ik} \mid O, I_i = 0)] = N_i^{-1} \sum_{k=1}^{N_i} \frac{\sum_{l\in s}[(\pi_l^{-1} - 1)K_l(x_{ik})]}{\sum_{l\in s}(\pi_l^{-1} - 1)}, \qquad (4.4)$$

where $K_l(x) = E_U(y_{lk} \mid x_{lk}, (l,k) \in U) = E_s\left\{ \frac{E_{re}[\tilde{w}(y_{lk}, x_{lk}) y_{lk} \mid x_{lk} = x, u_l, (l,k) \in R]}{E_{re}[\tilde{w}(y_{lk}, x_{lk}) \mid x_{lk} = x, u_l, (l,k) \in R]} \middle| O \right\}$.

See Pfeffermann and Sverchkov (2007, Section 7) for derivation of (4.4).

Remark 6. In a recent article, Verret et al. (2015**)** propose to account for informative sampling within the areas by including the sampling weights or functions of them as additional explanatory variables in the model. (The authors assume that all the areas are sampled with full response.) However, a similar approach cannot be used to account for NMAR nonresponse even with good estimates of the response probabilities since it requires knowledge of the area means of the probabilities $\tilde{\pi}_{k|i}$ for every area, but the response probabilities $\hat{p}_r(y_{ik}, x_{ik}; \gamma)$ can only be computed for the responding units.

5. MSE ESTIMATION

We propose a semi-parametric bootstrap procedure for MSE estimation. The procedure uses the same idea as in Sverchkov and Pfeffermann (2004). We first generate a pseudo population with marginal distributions of the outcome values, similar to the distributions of the true population values, and then select independently *B* samples from the pseudo population using the original sampling scheme and apply the same response mechanism as fitted to the original (true) sample. Finally, we compute the small area predictors for each area based on the sample of respondents.

A- Generation of pseudo population

P1. Use the observed data in order to regress the estimated area random effects and the area sampling weights, or functions of them, against area level variables such as $\bar{X}_i$ and N_i (and any other variables known at the area level), yielding the regression predictors, $w_i = g_w(\bar{X}_i, N_i)$ and $\hat{u}_i = g_u(\bar{X}_i, N_i)$. For non-sampled area k, set $\tilde{w}_k = \hat{g}_w(\bar{X}_k, N_k)$ and $\tilde{u}_k = \hat{g}_u(\bar{X}_k, N_k)$. For sampled area i, set $\tilde{u}_i = \hat{u}_i$ and $\tilde{w}_i = w_i$. Let $\tilde{\pi}_i = 1/\tilde{w}_i$.

P2. Generate a synthetic population with values $\tilde{y}_{ij} = \hat{p}_y(x_{ij}, \tilde{u}_i)$, $\tilde{\pi}_{j|i} = 1/\hat{\tilde{w}}(\tilde{y}_{ij}, x_{ij})$, ($\hat{\tilde{w}}(\tilde{y}_{ik}, x_{ik})$ is computed as below Eq. (4.2) with $\tilde{y}_{ik}$ instead of y_{ik}), and response probabilities $\tilde{p}_{ij}^{resp} = p_r(\tilde{y}_{ij}, x_{ij}; \hat{\gamma})$, $i = 1,...,M$, $j = 1,...,N_i$. Note that the synthetic population contains the same auxiliary variables as the original population and that the outcomes are generated from the model fitted to the *responding units*, but with estimated random effects and model coefficients.

P3. For each area $i = 1,...,\mathrm{M}$ of size N_i in the synthetic population, sample *with replacement* N_i units with probabilities proportional to $1/(\tilde{\pi}_{j|i}\tilde{p}_{ij}^{resp})$.

This concludes the generation of the pseudo-population.

Remark 7. As implied by the results of Sverchkov and Pfeffermann (2004), if the model hyper-parameters and random effects were actually known, the marginal distributions of the outcomes $\tilde{y}_{ij}$ in the pseudo-population would have been the same as the corresponding marginal distributions of the outcomes y_{ij} in the original population. In

practice, one can only use estimated parameters but as our simulation study shows, the procedure proposed in this section for MSE estimation, which relies on generating the pseudo population performs well even for areas with small samples. Notice in this respect that the model hyper-parameters are estimated from all the sampled areas, but for given hyper-parameter estimates, the estimates of the random effects are 'direct'.

B- Selection of bootstrap samples and computation of estimates

B1. Sample independently B samples $(\tilde{y}_{ij}^b, x_{ij}^b, \tilde{\pi}_{j|i}^b, \tilde{\pi}_i^b)$, $b = 1,...,B$; $i = 1...m$, $j = 1...n_i$, from the pseudo population using the same sampling schemes as used for selecting the original sample, but with inclusion probabilities $\tilde{\pi}_i$, $\tilde{\pi}_{j|i}$.

B2. For each unit in the sample initiate response with probability $p_r(\tilde{y}_{ij}^b, x_{ij}^b; \hat{\gamma})$, where $\hat{\gamma}$ is the estimate obtained from the true original sample.

B3. For each bootstrap sample *b*, re-estimate all the parameters of interest (means or totals in the present paper).

B4. Calculate empirical MSE or other statistics of interest over the *B* bootstrap samples.

As implied by the description of the proposed bootstrap method, we account for all the random processes underlying the population model, the informative sampling of areas and within the areas, and the response process.

6. SIMULATION STUDY

In this section we describe the results of a simulation study when applying the procedures proposed in Sections 3-5. In Section 7 we apply the method to a real data set.

6.1 Simulation set-up

The simulation study consists of the following 6 steps:

S1- *Generation of population values*: generate binary covariate values with $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0) = 0.5$, and corresponding outcome values from the mixed logistic model**,**

$$\Pr(y_{ij} = 1 \mid x_{ij}, u_i^U) = p_y(x_{ij}, u_i^U) = \frac{\exp(-0.1 - x_{ij} + u_i^U)}{1 + \exp(-0.1 - x_{ij} + u_i^U)}; \quad u_i^U \sim N(0, \sigma_u^2), \quad (6.1)$$

$i = 1,..,300$, $N_i = \text{int}[1000 \times \exp\{\min[2.5, \max(-2.5, u_i^U)]/5\}]$. The use of this function truncates the area size when the random effect is too small or too large.

We consider four different variances, $\sigma_u^2 = 1$, $\sigma_u^2 = 0.25$, $\sigma_u^2 = 0.1$ and $\sigma_u^2 = 0.01$ so as to study the effect of the magnitude of the variance on the performance of alternative estimators (see below).

Group the areas randomly into 3 sets,

G1={*i*=1,..,100}, G2={*i*=101,..,200}, G3={*i*=201,..,300}.

S2- *Sample selection*: select 50 areas from each group by systematic probability proportional to size (PPS) sampling with the area sizes, N_i, as the size variable. Notice that this implies an informative sampling of the areas since the size N_i depends on the random effect u_i^U. Select 20 units from each selected area in G1, 40 units from each selected area in G2 and 60 units from each selected area in G3 using systematic PPS sampling, with the size variable defined as, $z_{ij} = 5 + x_{ij} + 3y_{ij}$. This sampling scheme implies informative sampling of units within the selected areas since the size z_{ij} depends on the outcome y_{ij}.

S3- *Response mechanism*: obtain response from unit j in sampled area i with probability,

$$p_r(y_{ij}, x_{ij}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}, \qquad (6.2)$$

where $\gamma_0 = 0$, $\gamma_1 = -0.5$, $\gamma_2 = 2$. The nonresponse is NMAR since the response probability depends on the outcome. With these response probabilities the response rates are about 60%. We considered also a case where the response probabilities were generated from a different logistic model (Eq. 6.8 below). For this case the response rate was only 46%.

S4- *Fitting of respondent' model*: estimate, $\hat{p}_y(x_{ij}, u_i) = \hat{\Pr}(y_{ij} = 1 | x_{ij}, u_i, (i, j) \in R)$ by fitting the mixed logistic model,

$$p_y(x_{ij},u_i)=\frac{\exp(\beta_0+\beta_1 x_{ij}+u_i)}{1+\exp(\beta_0+\beta_1 x_{ij}+u_i)};\ \ u_i\sim N(0,\sigma_u^2), \qquad (6.3)$$

using PROC NLMIX in SAS with default options. Notice that the model (6.3) is not the true respondents' model under the population model (6.1), the informative sampling scheme described above and the response model (6.2).

S5. *Estimation of response probabilities*: assume, $p_r(y_{ij},x_{ij},\gamma)=\frac{\exp(\gamma_0+\gamma_1 x_{ij}+\gamma_2 y_{ij})}{1+\exp(\gamma_0+\gamma_1 x_{ij}+\gamma_2 y_{ij})}$, compute the expectations in (3.2) under the estimated model $\hat{p}_y(x_{ij},\hat{u}_i)$ in (6.3) and solve the resulting equations to estimate γ, using the procedure described in Section 3.

S6- *Prediction of area means*: first estimate $\tilde{w}(y_{ij},x_{ij}) = E_{re}[\tilde{\pi}_{j|i}^{-1}\mid y_{ij},x_{ij},(i,j)\in R]$ as follows: By definition, $E_{re}[\tilde{\pi}_{j|i}^{-1}\mid y_{ij},x_{ij},(i,j)\in R]=p_r^{-1}(y_{ij},x_{ij})E_{re}[\pi_{j|i}^{-1}\mid y_{ij},x_{ij},(i,j)\in R]$, where $\pi_{j|i}=n_i z_{ij}/\sum_{j=1}^{N_i} z_{ij}=\frac{n_i}{N_i}z_{ij}(1/\bar{Z}_i)$ and $\bar{Z}_i=\frac{1}{N_i}\sum_{j=1}^{N_i} z_{ij}$ is the i^{th} area mean, which is viewed as a constant, assuming that the true area size is large. Let $z_{ij}^{*}=\pi_{j|i}\frac{N_i}{n_i}$. In the present simulation study we fit the model $z_{ij}^{*}=\mathrm{g}_\alpha(y_{ij},x_{ij})=\alpha_0+\alpha_y y_{ij}+\alpha_x x_{ij}+\varepsilon_{ij}$, but other models can be fit, depending on the available data. Notice that $\mathrm{g}_\alpha(y_{ij},x_{ij})$ refers to the sample data before response and therefore the response weights $p_r^{-1}(y_{ij},x_{ij};\hat{\gamma})$ have to be used for estimating this model via weighted regression. Alternatively, one can fit the model for $p_r^{-1}(y_{ij},x_{ij};\hat{\gamma})\pi_{j|i}$ as a function of (y_{ij},x_{ij}), using the observed data (without weighting).

Estimate $\hat{\tilde{w}}(y_{ij},x_{ij})$ as,

$$\hat{\tilde{w}}(y_{ij},x_{ij})=\hat{E}_{re}[\tilde{\pi}_{j|i}^{-1}\mid y_{ij},x_{ij},(i,j)\in R]=\left[\frac{n_i}{N_i}\mathrm{g}_{\hat{\alpha}}(y_{ij},x_{ij})\right]^{-1}p_r^{-1}(y_{ij},x_{ij};\hat{\gamma}). \qquad (6.4)$$

Next, compute the ratios of the estimated expectations

$$\tilde{R}a(x_{ik},u_i)=\frac{E_{re}[(\hat{\tilde{w}}(y_{ik},x_{ik})-1)y_{ik}\mid x_{ik},u_i,(i,k)\in R]}{E_{re}[(\hat{\tilde{w}}(y_{ik},x_{ik})-1)\mid x_{ik},u_i,(i,k)\in R]} \qquad (6.5)$$

for estimating the conditional expectation of the missing outcomes in sampled areas. (The expectations in the ratio are computed similarly to the computation of the expectations in the example of section 3).

Finally, substitute (6.5) in (3.2) and estimate the mean outcome of sampled areas by substituting $\hat{u}_i$ for u_i and dropping the external expectation operator over the distribution of the random effects.

Remark 8. To save space, we only consider the prediction of the mean outcome in sampled areas, which are subject to NMAR nonresponse. The estimation of the means of non-sampled areas is the same as in Pfeffermann and Sverchkov (2007) and is illustrated in the simulation study of that paper.

Repeat Steps S1-S6 independently 500 times. The values x_{ij} of the covariate are generated only once and held fixed for all the simulations.

Predictors considered: compute the following predictors for each simulation.

1. $\hat{\bar{Y}}_i^{ign} = N_i^{-1}\{\sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} \hat{p}_y(x_{ij},u_i)\}$ with $\hat{p}_y(x_{ij},u_i) = p_y(x_{ij},\hat{u}_i)$; this estimator ignores the sampling and response process and "assumes" that the population distribution holds also for the respondents.

2. $\hat{\bar{Y}}_i^{HB,MCAR} = \sum_{j,(i,j)\in R} \pi_{j|i}^{-1} y_{ij} \,/\, \sum_{j,(i,j)\in R} \pi_{j|i}^{-1}$; this is the familiar Hajek-Brewer (Hajek, 1971) estimator that "assumes" that the nonresponse is completely at random.

3. $\hat{\bar{Y}}_i^{MAR} = \sum_{j,(i,j)\in R} \hat{w}(x_{ij}) y_{ij} \,/\, \sum_{j,(i,j)\in R} \hat{w}(x_{ij})$; $\hat{w}(x_{ij}) = [\pi_{j|i} p(x_{ij},\hat{\lambda})]^{-1}$; this estimator accounts for the response process but assumes that the nonresponse is MAR, and hence the response probabilities are estimated by assuming the propensity scores model $\Pr(R_{ij}=1 \mid x_{ij};\lambda) = p_r(x_{ij},\lambda) = \dfrac{\exp(\lambda_0+\lambda_1 x_{ij})}{1+\exp(\lambda_0+\lambda_1 x_{ij})}$. The parameter $\lambda = (\lambda_0,\lambda_1)'$ is estimated by solving the likelihood equations $\sum_{(i,j)\in R} \dfrac{\partial \log p_r(x_{ij};\lambda)}{\partial\lambda} + \sum_{(i,j)\in R^c} \dfrac{\partial \log[1-p_r(x_{ij};\lambda)]}{\partial\lambda} = 0$.

4. $\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i})$; already defined in (4.3). Accounts for NMAR nonresponse.

5. $\hat{\bar{Y}}_i^{new} = N_i^{-1}[\sum_{j,(i,j)\in R} y_{ij} + \sum_{k=1,k\notin R}^{N_i} \hat{R}a_{ij}]$; proposed empirical model-dependent predictor obtained from (4.2). The ratios $\hat{R}a_{ij}$ are obtained from (6.5) by substituting $\hat{u}_i$ for u_i.

The last two estimators are of prime interest as they account for both the informative sampling and NMAR nonresponse.

Statistics considered for assessment of performance of predictors and root MSE estimates

1- Prediction of area means: Let $D_{ir} = 1(0)$ if area *i* is sampled (not sampled) on the *r*-th simulation. Denote by $\bar{Y}_{ir}$ the true area mean of area *i* on the *r*-th simulation and let $\hat{\bar{Y}}_{ir}$ represent any of the five predictors defined above, $r = 1,...,500$.

$$Bias_i = \frac{\sum_{r=1}^{500} D_{ir}(\hat{\bar{Y}}_{ir} - \bar{Y}_{ir})}{\sum_{r=1}^{500} D_{ir}} \;;\; RMSE_i = \sqrt{\frac{\sum_{r=1}^{500} D_{ir}(\hat{\bar{Y}}_{ir} - \bar{Y}_{ir})^2}{\sum_{r=1}^{500} D_{ir}}}. \qquad (6.6)$$

2- Estimation of Root MSE (RMSE): Because of running time limitations, for estimation of the RMSE we only considered the first 100 simulations and generated only 50 bootstrap samples for each simulation. Let $D_{irb} = 1(0)$ if area *i* is sampled (not sampled) in the *b*-th bootstrap sample on the *r*-th simulation. Denote by $\bar{Y}_{ipr}$ the pseudo area mean of area *i* on the *r*-th simulation and let $\hat{\bar{Y}}_{irb}^{new}$ represent the corresponding new predictor.

$$RMSE_i^{Boot} = \sqrt{\frac{1}{100}\sum_{r=1}^{100} \hat{MSE}_{i,r}^{Boot}} \;;\; \hat{MSE}_{i,r}^{Boot} = \frac{\sum_{b=1}^{50} D_{irb}(\hat{\bar{Y}}_{irb}^{new} - \bar{Y}_{ipr})^2}{\sum_{b=1}^{50} D_{irb}}. \qquad (6.7)$$

In any given application, one would obviously generate many more bootstrap samples but notice that we report summary statistics over the 100 simulations, so we actually report the results obtained over $100 \times B_i$ bootstrap samples, where B_i is the number of times that area i has been sampled.

6.2 Results for the case of "large" random effects ($\sigma_u^2 = 1$)

In this section we consider the case of relatively "large" random effects.

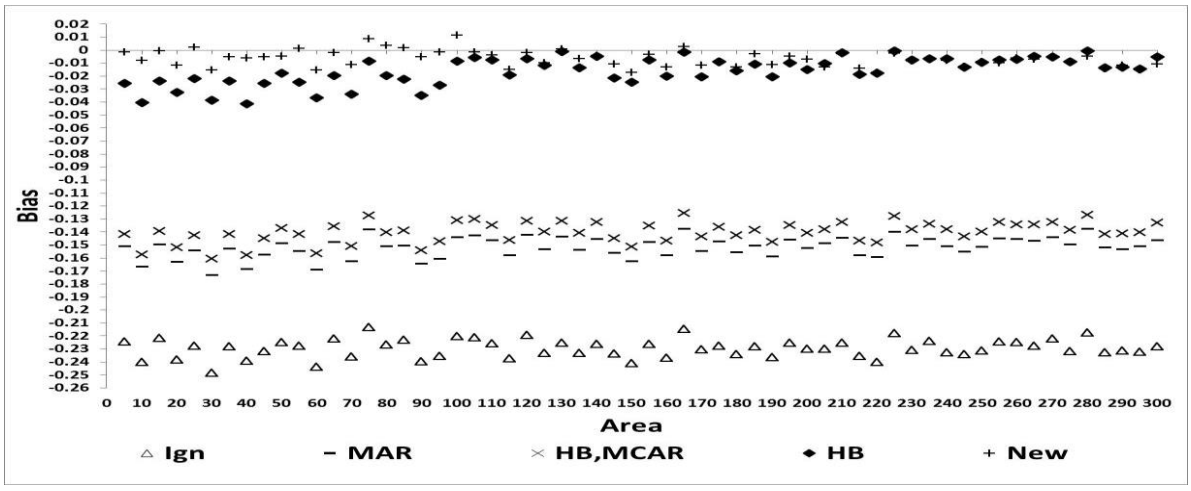
Table 1. Estimation of response model coefficients

	$\gamma_0 = 0$	$\gamma_1 = -0.5$	$\gamma_2 = 2$
Bias	0.006	0.003	0.037
Std	0.055	0.045	0.174

Although the estimators of all three coefficients are biased, the biases are relatively very small and so are the standard deviations (Std). The small biases have negligible effect on the estimation of the true response probabilities. The mean of the true response probabilities over the 500 simulations turned out to be 0.625, and the mean of the corresponding estimated probabilities is 0.624. The mean over the 500 simulations of the standard deviations of the differences between the true and the corresponding estimated probabilities is 0.012.

The figures that follow illustrate the performance of the procedure at the area level. To make the figures clearer, we ordered the areas in each of the three groups according to their size, N_i, and we show the results for every 5th area.

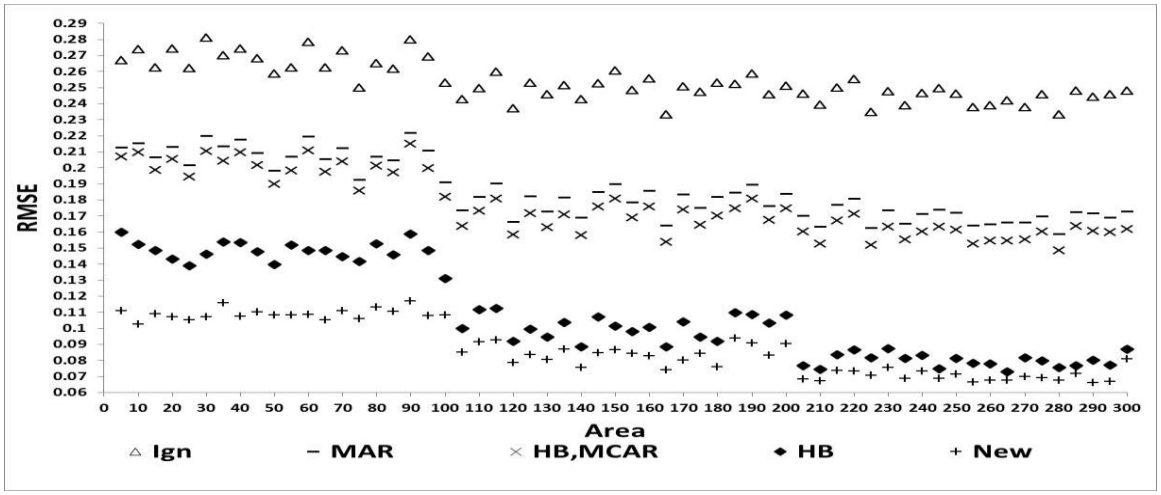
Figure 1. Bias of predictors by area, 500 simulations



The conclusions from Figure 1 are clear-cut. The proposed model-dependent predictor $\hat{\bar{Y}}_i^{new}$ is virtually unbiased for each of the areas. The Hajek-Brewer estimator is also nearly unbiased, except in the areas with the small sample sizes. (Despite using estimated probabilities, it is a ratio-type estimator). The other three predictors, which ignore the

informative response process are biased, with particularly large bias of the predictor $\hat{\bar{Y}}_i^{ign}$ that ignores both the informative sampling and the response.

Figure 2. RMSE of predictors by area, 500 simulations



The RMSE of our proposed predictor, $\hat{\bar{Y}}_i^{new}$ is uniformly the smallest, with the Hajek-Brewer estimator being second in order. The RMSE of $\hat{\bar{Y}}_i^{ign}$ is dominated by its large bias and hence its large value. The RMSE of all the predictors decrease as the sample sizes increase, due to decrease in the variance.

Figure 3. Estimation of RMSE of $\hat{\bar{Y}}_i^{new}$ by area

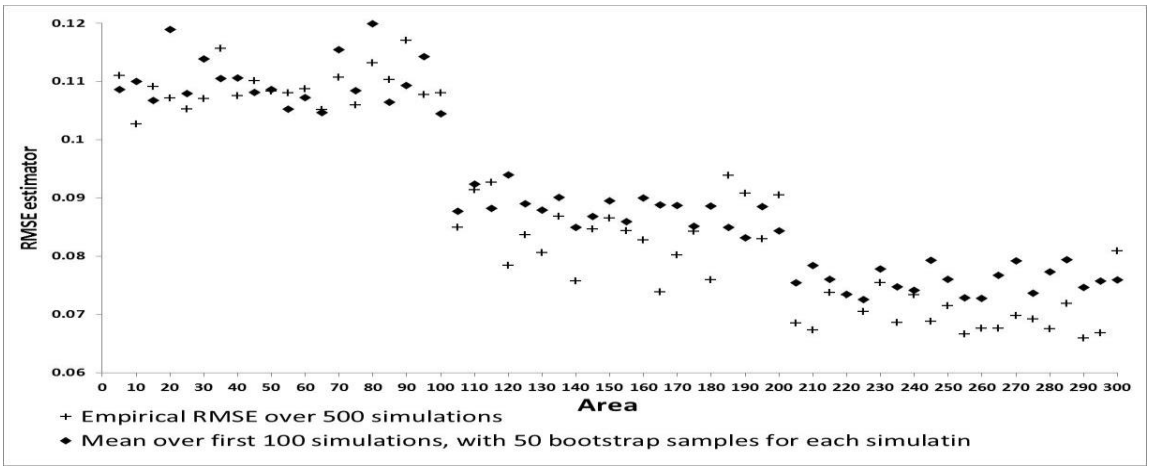


Figure 3 indicates good performance of the bootstrap RMSE estimates in terms of bias.

Next we illustrate the robustness of the model assumed for the response probabilities, discussed in *Remark* 4. For this, we repeated the same simulation study but with a different true response model,

$$p(y_{ij}, x_{ij}, \gamma) = \frac{\exp(-.5x_{ij}y_{ij})}{1+\exp(-.5x_{ij}y_{ij})}. \tag{6.8}$$

(Compare with (5.2)). However, we fit the response model

$p_r(y_{ij}, x_{ij}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}{1+\exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})}$ (same as before. We did not add the cross product $x_{ij}y_{ij}$ to the model because it would make the true response model a special case and we want to illustrate the robustness of the working response model. Notice also that X and Y are binary, so that there is no point of adding polynomials of these variables.)

In this case there is nothing to compare the estimated response model coefficients with, but we can still compare the true response probabilities with the estimated probabilities. The mean of the true response probabilities over the 500 simulations is now 0.457 and the mean of the estimated probabilities is 0.456. The mean over the 500 simulations of the standard deviations of the differences between the true and the corresponding estimated probabilities is 0.06. Thus, even though the response model is strongly misspecified, the estimation of the response probabilities is still unbiased, although with greater variability. (The mean of the standard deviations was 0.012 when estimating the correct response model.) As shown in the next three figures, the prediction of the true area means is likewise reliable and much better than when ignoring the NMAR response process.

Figure 4. Bias of predictors by area, response model misspecified, 500 simulations

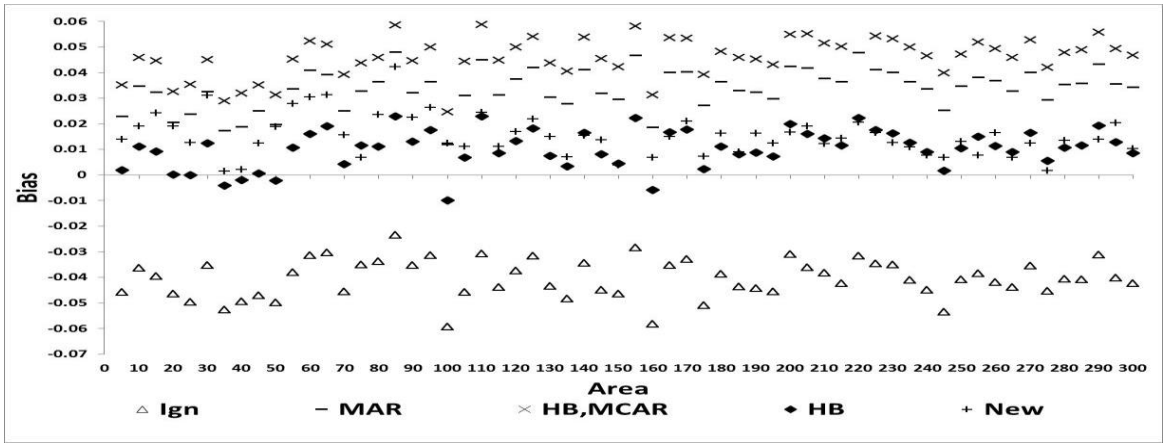
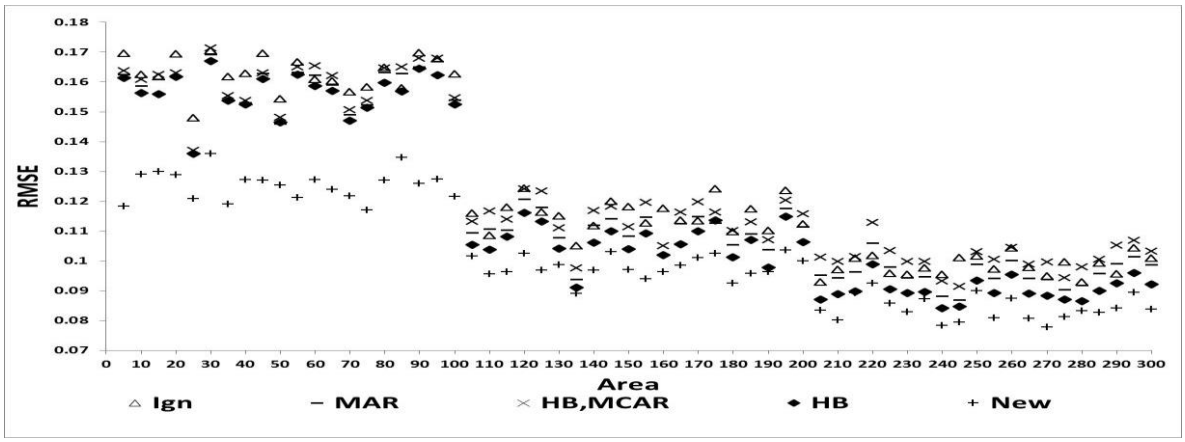


Figure 4 exhibits a similar picture to Figure 1, with the new and the Hajek-Brewer predictors being now slightly biased. The other three predictors are more biased, but the bias is considerably smaller than in Figure 1, as obtained when estimating the correct response model. The different magnitudes of the bias of the three predictors in Figures 1 and 4 is explained by the fact that since the response model is different in the two cases, so is the respondents' distribution, resulting in different distributions of the estimators that ignore the informative sampling or nonresponse. Thus, Figures 1 and 4 are not really comparable.

Figure 5. RMSE of predictors by area, response model misspecified, 500 simulations



The RMSEs of the proposed- and the Hajek-Brewer predictors change only slightly when misspecifying the model of the response probabilities. The RMSEs of the other three predictors are smaller under the misspecified model, due to the decrease in the bias.

Figure 6. Estimation of RMSE of $\hat{\bar{Y}}_i^{new}$ by area, response model misspecified

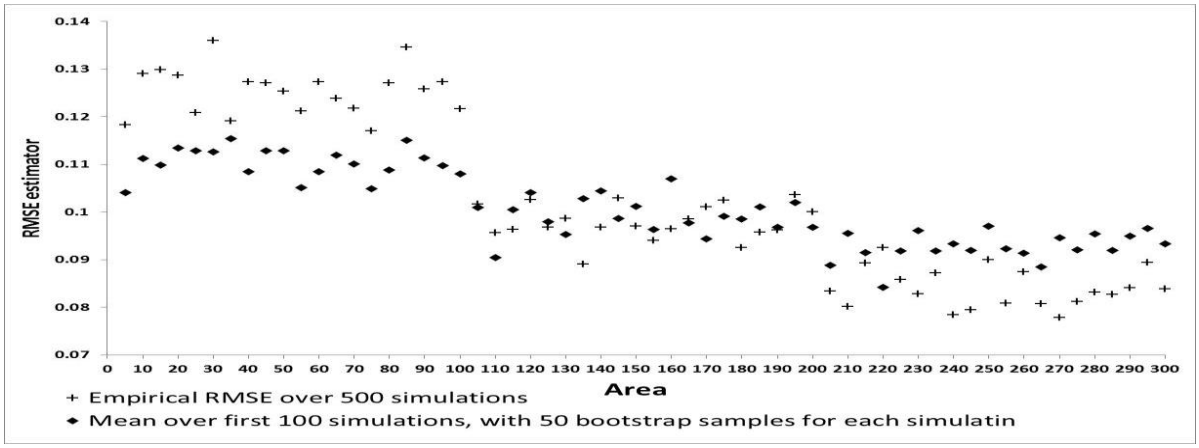


Figure 6 indicates a negative bias of the RMSE estimators in the areas with small sample sizes, which decreases in absolute value as the sample size increases.

All in all, this part of the simulation study supports the discussion in Remark 4 regarding the robustness of the proposed procedure with a logistic response model to possible misspecifications of this model.

6.3 Results for "medium size" random effects ($\sigma_u^2 = 0.25$)

In this section we consider the case where the random effects are of much lower magnitude, as defined by their variance. The results in this section are again based on 500 simulations.

Table 2. Estimation of response model coefficients

	$\gamma_0 = 0$	$\gamma_1 = -0.5$	$\gamma_2 = 2$
Bias	-0.093	0.042	0.288
Std	0.102	0.063	0.308

As expected, the bias of all the three estimators are now larger, and so are the standard deviations, but the biases are still relatively small and as illustrated below, have little effect on the estimation of the response probabilities. The mean of the true response probabilities over the 500 simulations is in this case 0.623 and the mean of the estimated response probabilities is 0.617. The mean over the 500 simulations of the standard deviations of the differences between the true and the corresponding estimated probabilities is 0.029 (compared to 0.012 when $\sigma_u^2 = 1$). Notice that decreasing the variance of the random effects does not make the response probabilities and sample selection probabilities less informative. For example, for $\sigma_u^2 = 1$, the average of the response probabilities was found to be 0.625, with average standard deviation of 0.220. (We first computed the average and standard deviation for each simulation and then averaged them over the 500 simulations.) The corresponding figures for the within area sample selection probabilities are 0.0196 and 0.00557. For $\sigma_u^2 = 0.25$, the average of the response probabilities was found to be 0.623 with average standard deviation of 0.221.

The corresponding figures for the within area sample selection probabilities are 0.0196 and 0.00572.

As in Section 6.2, the figures that follow illustrate the performance of the various predictors at the area level.

Figure 7. Bias of predictors by area, 500 simulations

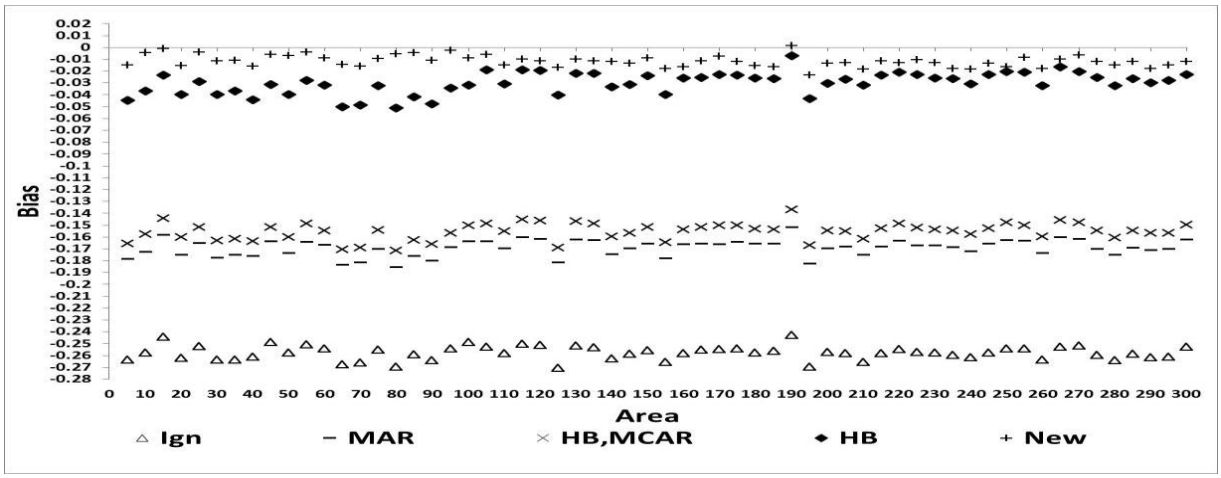


Figure 8. RMSE of predictors by area, 500 simulations

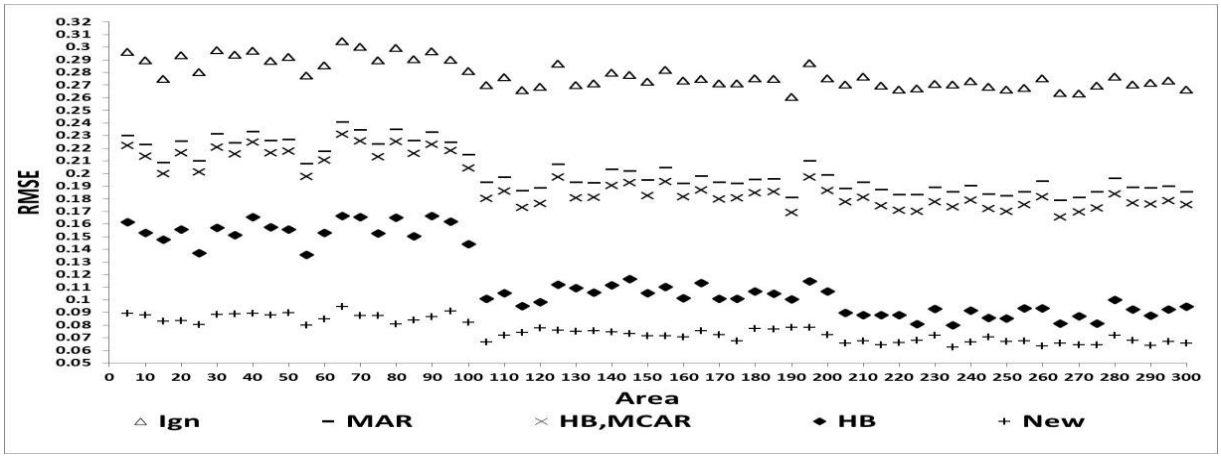
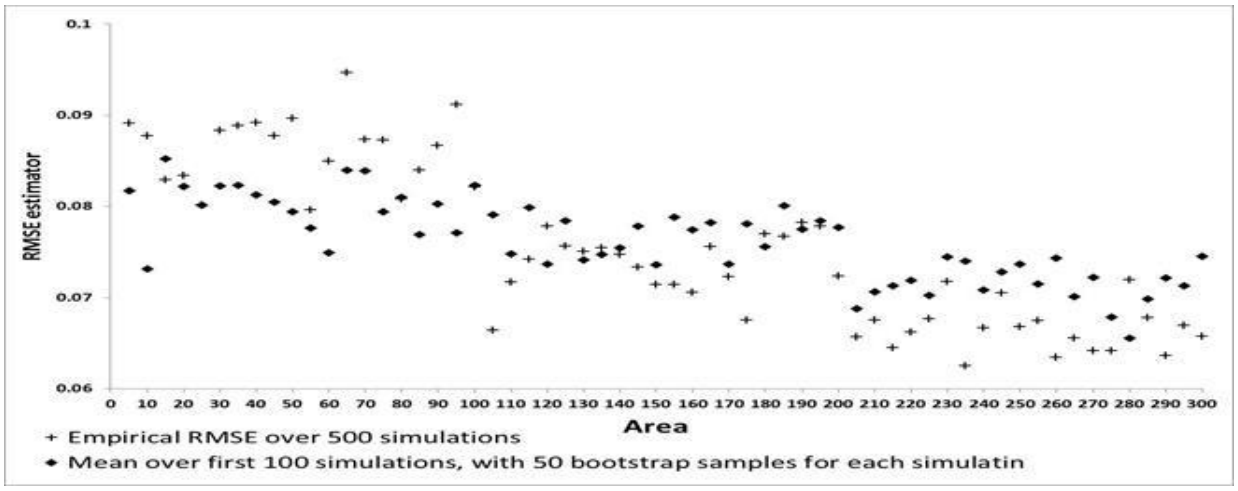


Figure 9. Estimation of RMSE of $\hat{\bar{Y}}_i^{new}$ by area



The general conclusion from Table 2 and Figures 6-9 is that the proposed procedure (the predictors $\hat{\bar{Y}}_i^{HB}$ and $\hat{\bar{Y}}_i^{new}$) works well in removing the bias resulting from informative sampling and NMAR nonresponse, also with the much smaller variance of $\sigma_u^2 = 0.25$.

Remark 9. We repeated the simulation study also for the case $\sigma_u^2 = 0.1$ and the two predictors $\hat{\bar{Y}}_i^{HB}$ and $\hat{\bar{Y}}_i^{new}$ still perform much better than the other three predictors considered, although they now have a somewhat larger bias and RMSE. However, this is no longer the case when $\sigma_u^2 = 0.01$, in which case all the five predictors perform badly because of failure to estimate properly the response probabilities, in line with the discussion in *Remark 3*. This implies an interesting contrast because in SAE models, one usually attempts to include in the model as many covariates as possible, so as to reduce the unexplained variations represented by the random effects (small σ_u^2). However, if no other covariates x_2 that explain the outcome variable but not the response exist, then it is important that the variance σ_u^2 of the random effects is not too small, thus allowing to estimate the response probabilities and remove the bias induced by NMAR nonresponse.

7. PREDICTION OF NUMBER OF MARRIED PEOPLE IN SMALL STATISTICAL AREAS IN ISRAEL

7.1 Motivation and Background

Israel has a fairly accurate population register. In fact, at the country level, the register is almost perfect, because of accurate records of births, deaths and immigrants. The only real problem, shared by other countries, is the enumeration of emigrants, as it is hard to define emigrants and count them. However, population counts are required for small domains, as defined by 'statistical areas', with an average size of about 3,000 persons. For these small domains, the population register is much less accurate, with an average enumeration error of about 13 percent and a 95th percentile of 40 percent. The main reason for the inaccuracy of the register at the statistical area level is that people moving in or out an area are often slow to report their change of address. This occurs mostly among young adults who tend to change addresses more frequently because of change of jobs, school catchment areas of their children, and/or differences in house values, rental prices and municipal tax rates between geographic regions.

To deal with this problem, the Israel Central Bureau of Statistics (ICBS) conducted in 2008 an integrated (dual system) census, which consisted of the population register, corrected by estimates obtained from two *coverage samples* for each statistical area: an area sample of addresses for estimating the register *undercount* (people living in the area but not registered there), and a telephone sample of people registered in the area for estimating the *register overcount* (people registered falsely as living in the area). Denote,

N_i - true number of people living in area i,

K_i - number of people registered as living in area i,

$p_{i,L|R}$ - proportion of people living in area i among those registered in the area

$p_{i,R|L}$ - proportion of people registered to area i among those living in the area.

Then,

$$N_i \times p_{i,R|L} = K_i \times p_{i,L|R} \Leftrightarrow N_i = K_i \times \frac{p_{i,L|R}}{p_{i,R|L}}. \qquad (7.1)$$

Thus, N_i is estimated from the two samples as,

$$\hat{N}_i = K_i \times \frac{\hat{p}_{i,L|R}}{\hat{p}_{i,R|L}}, \tag{7.2}$$

where $\hat{P}_{i,R|L}$ and $\hat{P}_{i,L|R}$ are the corresponding design-based estimators from the two samples. The design variance of $\hat{N}_i$ can be approximated by Taylor linearization as,

$$Var(\hat{N}_i \mid K_i) = K_i^2 \left[\frac{Var(\hat{p}_{i,L|R})}{[E(\hat{p}_{i,R|L})]^2} + \frac{[E(\hat{p}_{i,L|R})]^2}{[E(\hat{p}_{i,R|L})]^4} \times Var(\hat{p}_{i,R|L}) \right]. \tag{7.3}$$

In what follows we restrict to the overcount survey. Prior to the phone calls, a letter was sent to all the sampled members notifying them of the survey and asking them to respond to the phone interview. Nonetheless, there is a high rate of nonresponse in this survey, with an average response rate of about 0.75 and standard deviation between areas of about 0.14. Moreover, it is quite obvious that the nonresponse is NMAR because the nonrespondents are more likely to be the persons not registered correctly (living in another area) and hence not getting the notice letter in the first place.

The sampling design used in each statistical area is systematic sampling after ordering the frame by age. Notice that this sampling scheme is *noninformative* since all the sampling units in a given area have the same sampling probability. The target is to estimate the total number of persons registered as living in the area and actually living there. Ideally, we would have wanted to show how our proposed procedure performs in reducing the bias of the naïve estimates, which ignore the nonresponse altogether. However, we have no information on the true target numbers of people registered correctly, so that analyzing this data set would not allow us to draw any conclusions. Consequently, we show below the performance of the various predictors when predicting the number of married people registered as living in the area, which is known to be correlated with correct registration. The true counts are known for every area from the population register from which the sample is taken.

Let $y_{ij} = 1$ if person j registered as living in area i is married and $y_{ij} = 0$ otherwise. Let x_{ij} denote the age of person (i, j) and define $X_{1ij} = 1$ if $x_{ij} > 25$, $X_{1ij} = 0$ otherwise; $X_{2ij} = 1$ if $x_{ij} > 40$, $X_{2ij} = 0$ otherwise. The following logistic models have been

assumed for the outcomes observed for the responding units, and for the response probabilities, after removing from the data set persons aged 16 or less, which are all singles. These persons have been added back when computing the final estimates.

$$p_y(\mathrm{x}_{ij}, u_i) = \frac{\exp(\beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + u_i)}{1 + \exp(\beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + u_i)}; \;\; u_i \sim N(0, \sigma_u^2), \tag{7.4}$$

$$p_r(y_{ij}, \mathrm{x}_{ij}; \gamma) = \frac{\exp(\gamma_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \gamma_3 y_{ij})}{1 + \exp(\gamma_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \gamma_3 y_{ij})}. \tag{7.5}$$

Denote by R_i the subsample of responding persons in the sample of persons registered as residing in area i. We computed for every area i the following predictors of the number of married people, M_i, among the K_i persons registered as living in the area.

$\hat{M}_{i,MCR} = K_i \times (\sum_{j \in R_i} y_{ij} \,/\, \sum_{j \in R_i} 1)$; assumes MCAR nonresponse

$$\hat{M}_{i,MR} = \left(\sum_{j \in R_i} y_{ij} + \sum_{k=1, k \notin R_i}^{N_i} \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1ik} + \hat{\beta}_2 X_{2ik} + \hat{u}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1ik} + \hat{\beta}_2 X_{2ik} + \hat{u}_i)} \right)$$; assumes MAR nonresponse

$$\hat{M}_i^{HB} = K_i \times \frac{\sum_{j \in R_i} y_{ij} \times \hat{p}_r^{-1}(y_{ij}, \mathrm{x}_{ij}; \gamma)}{\sum_{j \in R_i} \hat{p}_r^{-1}(y_{ij}, \mathrm{x}_{ij}; \gamma)}$$; Hajek-Brewer estimator with estimated probabilities.

$\hat{M}_i^{new} = \sum_{j \in R_i} y_{ij} + \sum_{k=1, k \notin R_i}^{N_i} \hat{R} a_{ik}$; proposed predictor. See Section 6.1.

7.2 Results

We consider separately areas of size $N_i \leq 100$, (A=565 areas, with an average size of 75 persons), and areas of size $100 < N_i \leq 200$, (A=370 areas, with an average size of 132 persons) the small population areas. The mean number of responding units in the first group of areas is 63.48, with standard deviation of 27.92. The corresponding figures in the second group are 128.67 and 24.7. Denote by $E_i = (M_i - \hat{M}_i)$ the prediction error, where $\hat{M}_i$ represents any of the 4 predictors. Table 3 contains the following summary statistics for the four predictors over all the areas in each of the two groups.

$$Bias_{Err} = \sum_{i=1}^{A} E_i / A \;;\;\; RMSE_{Err} = (\sum_{i=1}^{A} E_i^2 / A)^{0.5}. \qquad (7.6)$$

Table 3. Bias and RMSE of prediction errors over all the areas in each group

Predictors		$\hat{M}_{i,MCAR}$	$\hat{M}_{i,MAR}$	$\hat{M}_{i,HB}$	$\hat{M}_{i,NEW}$
Small areas	$Bias_{Err}$	2.95	2.25	0.88	0.79
	$RMSE_{Err}$	4.45	3.26	3.28	2.18
Larger areas	$Bias_{Err}$	4.50	3.32	0.68	0.67
	$RMSE_{Err}$	5.86	4.52	4.02	3.05

Small areas ($N_i \leq 100$) - $\hat{\gamma}_0 = 0.56, \hat{\gamma}_y = 0.67, \hat{\gamma}_{X1} = 0.23, \hat{\gamma}_{X2} = 0.22$, $\hat{\sigma}_u^2$ **= 0.29**
True mean number of married people= **24.40**; standard deviation (between areas)= **13.25**

Larger areas ($100 < N_i \leq 200$) - $\hat{\gamma}_0 = 0.78, \hat{\gamma}_y = 0.95, \hat{\gamma}_{X1} = 0.08, \hat{\gamma}_{X2} = 0.25$, $\hat{\sigma}_u^2$ **= 0.49**
True mean number of married people= **51.68**; standard deviation (between areas)= **11.13**

The results in Table 3 indicate very clearly that the two predictors that account for NMAR nonresponse perform much better than the other two predictors. Among the two, the proposed predictor, $\hat{M}_{i,NEW}$ has a smaller RMSE, as is the case also in the simulation study (Figures 2, 5 and 8). Notice the relatively large values of the coefficients $\hat{\gamma}_y$ in the two response models, indicating a high degree of informativeness of the nonresponse. Also notice how the bias of the two predictors is reduced, as the estimated variance of the random effects increases from $\hat{\sigma}_u^2 = 0.29$ to $\hat{\sigma}_u^2 = 0.49$.

8. SUMMARY

In this article we propose a general approach for small area estimation under informative sampling of areas and within areas, and NMAR nonresponse within the selected areas. The approach consists of identifying a model holding for the observed data with non-negligible random effects (as is usually the case with small area models), and using this model for estimating the response probabilities by application of the Missing Information Principle. The use of this principle assumes a parametric model for the response probabilities as a function of the covariates and the outcome, but we review theoretical results justifying the use of a logistic model with appropriate powers and interactions of

the outcome and the covariates as a good approximation to the true response mechanism. Once the response probabilities are estimated, we consider them as known and follow the approach of Pfeffermann and Sverchkov (2007) for estimating the area means under informative sampling (assuming full response). We propose a bootstrap method for estimating the RMSE of the resulting predictors. WE also consider the much simpler Hajek-Brewer estimator as obtained by substituting the unknown response probabilities by their estimators. A simulation study shows good performance of the proposed approach and illustrates its robustness to misspecification of the response model. Application of the approach to a real data set further supports the use of this approach.

The empirical study in this article considers the case where the models fitted for the responding units and the response probabilities are logistic, but the theoretical derivations of our proposed approach assume general models for the observed data and the response mechanism. Thus, we encourage researchers of SAE to apply the procedure to other models fitted to the observed data, with possibly different sampling schemes and models assumed for the response probabilities.

As in Pfeffermann and Sverchkov (2007), the proposed methodology of the present article is under the frequentist approach. As is well known, there exists a vast literature on SAE under a full Hierarchical or empirical Bayes setting. Thus, an important intriguing challenge for future research would be to apply the proposed methodology in a Bayesian setup, with appropriate prior distributions for the models' hyper-parameters. See Pfeffermann et al. (2006) for application of the Bayesian approach for two-level modelling under informative sampling of first and second level units.

9. REFERENCES

Cepillini, R., Siniscialco, M., and Smith, C.A.B. (1955). The estimation of gene frequencies in a random mating population. *Annals of Human Genetics*, **20**, 97-115.

Feder, M., and Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and nonresponse- an empirical likelihood approach. Southampton Statistical Sciences Research Institute, University of Southampton, UK. http://eprints.soton.ac.uk/id/eprint/378245.

Hajek, J. (1971). Comments on paper by D. Basu. In: *Foundations of Statistical Inference* (eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston. P. 236.

Kim, J.K. and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, **100**, 385-398.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, **37**, 115-136.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40-68.

Pfeffermann, D., Moura F., and Silva, P. (2006). Multi-Level Modelling Under Informative Probability Sampling. *Biometrika*, 93, 943-959.

Pfeffermann, D. and Sikov N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, **27**, 181–209.

Pfeffermann, D., and Sverchkov, M. (2007). Small-Area Estimation under Informative Probability Sampling of Areas and Within Selected Areas. *Journal of the American Statistical Association*, **102**, 1427-1439.

Pfeffermann, D., and Sverchkov, M. (2009). Inference under Informative Sampling. In: Handbook of Statistics 29B; *Sample Surveys: Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 455-487.

Rao, J.N.K., and Molina, I. (2015), *Small Area Estimation*, 2nd Edition, Wiley.

Riddles, K.M., Kim, J.K. and Im, J. (2016). A propensity-scoreadjustment method for nonignorable nonresponse, Journal of Survey Statistics and Methodology, 4, 215-245.

Rivers, D. (2007). Sampling for web surveys. Joint Statistical Meeting, Proceedings of the Section on Survey Research Methods, Salt Lake City, UT, USA.
http://www.laits.utexas.edu/txp_media/html/poll/files/Rivers_matching.pdf

Särndal, C.E., and Swensson B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, **55**, 279-294.

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. *Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods*, 867-874.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of Finite Population Totals Based on the Sample Distribution. *Survey Methodology*, **30**, 79-92.

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