


# Small area estimation under informative sampling and not missing at random non-response

Michael Sverchkov

*Bureau of Labor Statistics, Washington DC, USA*

and Danny Pfeffermann

 *Central Bureau of Statistics, Jerusalem, Hebrew University of Jerusalem, Israel, and University of Southampton, UK*

[Received May 2017. Revised January 2018]

**Summary.** Pfeffermann and Sverchkov considered small area estimation 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. We extend this approach to the practical situation of incomplete response at the unit level, and where the response is not missing at random. 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 Pfeffermann and Sverchkov 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 mean-squared error 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/or non-response.

**Keywords:** Missing information principle; Population distribution; Respondents model; Sample distribution

## 1. Introduction

Over the last 20 years, many references 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 in the sample 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 paper, not accounting for informative sampling or non-response can result in large bias and root-mean-squared error (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.

*Address for correspondence:* Danny Pfeffermann, Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.  
E-mail: msdanny@soton.ac.uk

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 paper, using an important result of Verret *et al.* (2015).

A related but definitely more complicated problem when analysing survey data is not missing at random (NMAR) non-response. 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 us to assume some structure for these probabilities. Because NMAR non-response 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, which is commonly known as missingness at random (MAR). It is far beyond the scope of this paper 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 paper is to propose a method of handling NMAR non-response in the framework of SAE. In official statistics, the sample that is used for SAE is basically the same sample as 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 non-response are the same in both cases, although the problems resulting from the non-response 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 reference has been published considering this very important problem of NMAR non-response. For this, we extend the approach of Pfeffermann and Sverchkov (2007). The extension consists of identifying the model holding for the observed responses and using this model for estimating the response probabilities by application of the missing information principle. For this, we define the likelihood holding for the sample under complete response, we integrate out the unobserved outcomes from this likelihood over the distribution holding for the non-respondents 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 mean-squared error (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 with predictors that ignore the informative sampling process or the NMAR non-

response mechanism. The procedure proposed is applied to a real data set from Israel in Section 7. We conclude with a brief summary in Section 8.

The program that was used for the simulation that produced the results in Table 1 and Figs 1–3 and the data for the figures can be obtained from

<http://wileyonlinelibrary.com/journal/rss-datasets>

## 2. Notation and models

Let  $\{y_{ij}, \mathbf{x}_{ij}; i = 1, \dots, M, j = 1, \dots, N_i\}$  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  $\mathbf{x}'_{ij} = (x_{ij,1}, \dots, x_{ij,K})$  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

$$\begin{aligned} y_{ij}|\mathbf{x}_{ij}, u_i^U &\sim f(y_{ij}|\mathbf{x}_{ij}, u_i^U), & i = 1, \dots, M, \quad j = 1, \dots, N_i, \\ u_i^U &\sim f(u_i^U), & E(u_i^U) = 0, \quad V(u_i^U) = \sigma_u^2, \end{aligned} \quad (2.1)$$

where  $u_i^U$  is the  $i$ th area level random effect under this model. The target is to estimate the area means  $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ ,  $i = 1, \dots, M$ , on the basis of a sample that is 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  $\pi_{j|i} = \Pr(j \in s_i | i \in s)$ . Denote by  $I_i$  and  $I_{ij}$  the sample indicators:  $I_i = 1$  if area  $i$  is selected in the first stage and  $I_i = 0$  otherwise;  $I_{ij} = 1$  if unit  $j$  of selected area  $i$  is sampled in the second stage and  $I_{ij} = 0$  otherwise. Let  $w_i = 1/\pi_i$  and  $w_{j|i} = 1/\pi_{j|i}$  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 non-respondents among the sampled units is  $R^c = \{(i, k) : I_i = 1, I_{ik} = 1, R_{ik} = 0\}$ . We assume that  $\sum_{j=1}^{n_i} R_{ij} > 0$  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,  $u_i = u_i^U - E(u_i^U | i \in s)$ . Then, under the population model (2.1), the observed data follow the two-level ‘respondents’ model

$$\begin{aligned} y_{ij}|\mathbf{x}_{ij}, u_i &\sim f_R(y_{ij}|\mathbf{x}_{ij}, u_i) = f\{y_{ij}|\mathbf{x}_{ij}, u_i, (i, j) \in R\}, \\ u_i &\sim f(u_i | i \in s), \quad E(u_i | i \in s) = 0. \end{aligned} \quad (2.2)$$

Model (2.2) is again general and all that we state at this stage is that, under informative sampling and/or NMAR non-response, the population and the respondents models differ:  $f_R(y_{ij}|\mathbf{x}_{ij}, u_i) \neq f(y_{ij}|\mathbf{x}_{ij}, u_i^U)$ .

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

Let  $p_r(y_{ij}, \mathbf{x}_{ij}) = \Pr(R_{ij} = 1 | y_{ij}, \mathbf{x}_{ij}, i \in s, j \in s_i)$ . If the probabilities  $p_r(y_{ij}, \mathbf{x}_{ij})$  were known, the sample of respondents could be considered as a two-stage sample from the finite population

with known selection probabilities  $\pi_i$  and  $\tilde{\pi}_{j|i} = \pi_{j|i} p_r(y_{ij}, \mathbf{x}_{ij})$ . 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):

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

Note that, under informative NMAR non-response, the non-respondents' distribution differs from the respondents' distribution  $f\{y_{ij}|\mathbf{x}_{ij}, u_i, (i, j) \in R\}$ . 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 non-respondents' distribution via equation (2.3). In the following section we show how we can estimate the response probabilities.

### 3. Estimation of response probabilities

Following Sverchkov (2008), we assume a parametric model for the response probabilities, which depends on an unknown vector parameter  $\gamma$ ;  $p_r(y_{ij}, \mathbf{x}_{ij}) = p_r(y_{ij}, \mathbf{x}_{ij}; \gamma) = \Pr(R_{ij} = 1|y_{ij}, \mathbf{x}_{ij}, i \in s, j \in s_i; \gamma)$ .

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

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

*Condition 1.*  $f\{y_{il}, y_{ij}|u_i, \mathbf{x}_{ij}, (i, l) \notin R, R_{ij} = 1\} = f\{y_{il}|u_i, \mathbf{x}_{il}, (i, l) \notin R\} f(y_{ij}|u_i, \mathbf{x}_{ij}, R_{ij} = 1)$ .

*Condition 2.*  $f\{\pi_{j|i}|u_i, y_{il}, y_{ij}, \mathbf{x}_{ij}, (i, l) \notin R, R_{ij} = 1\} = f(\pi_{j|i}|u_i, y_{ij}, \mathbf{x}_{ij}, R_{ij} = 1)$ .

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,  $\gamma$  could be estimated by solving the likelihood equations

$$\sum_{(i,j) \in R} \frac{\partial \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} + \sum_{(i,k) \in R^c} \frac{\partial \log\{1 - p_r(y_{ik}, \mathbf{x}_{ik}; \gamma)\}}{\partial \gamma} = 0. \quad (3.1)$$

In practice, the missing data are unobserved and hence the likelihood equations (3.1) are not operational. However, we may apply in this case the *missing information principle* (Cepillini *et al.*, 1955; Orchard and Woodbury, 1972): since no observations are available for  $(i, k) \in R^c$ , solve instead

$$\begin{aligned}
E_U & \left( \left[ \sum_{(i,j) \in R} \frac{\partial \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} + \sum_{(i,k) \in R^c} \frac{\partial \log\{1 - p_r(y_{ik}, \text{[redacted]})\}}{\partial \gamma} \right] \middle| O \right) \\
&= \sum_{(i,j) \in R} \frac{\partial \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} + \sum_{(i,k) \in R^c} E_{\text{nre}} \left[ \frac{\partial \log\{1 - p_r(y_{ik}, \text{[redacted]})\}}{\partial \gamma} \middle| O, (i,k) \in R^c \right] \\
&= \sum_{(i,j) \in R} \frac{\partial \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} \\
&\quad + \sum_{(i,k) \in R^c} E_s \left( E_{\text{nre}} \left[ \frac{\partial \log\{1 - p_r(y_{ik}, \text{[redacted]})\}}{\partial \gamma} \middle| O, u_i, (i,k) \in R^c \right] \middle| O, (i,k) \in R^c \right)
\end{aligned}$$

which equals, by equation (2.3) and assumption 2,

$$\begin{aligned}
& \sum_{(i,j) \in R} \frac{\partial \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} \\
&+ \sum_{(i,k) \in R^c} E_s \left( \frac{E_{\text{re}}[\{p_r^{-1}(y_{ik}, \text{[redacted]}) - 1\} \partial \log\{1 - p_r(y_{ik}, \text{[redacted]})\} / \partial \gamma \mid u_i, (i,k) \in R]}{E_{\text{re}}[\{p_r^{-1}(y_{ik}, \text{[redacted]}) - 1\} \mid u_i, (i,k) \in R]} \middle| O \right) \\
&= 0.
\end{aligned} \tag{3.2}$$

In equation (3.2)  $E_U$ ,  $E_s$ ,  $E_{\text{re}}$  and  $E_{\text{nre}}$  define respectively expectations with respect to the population distribution, the sample distribution, the respondents' distribution and the non-respondents' distribution. Note 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 missing information principle is simple. Ideally, we would want to use the score function (3.1) but, since the outcomes are unknown for the non-responding units, we replace the second expression,  $\sum_{(i,k) \in R^c} \partial \log\{1 - p_r(y_{ik}, \text{[redacted]})\} / \partial \gamma$  by its 'best predictor', as defined by its expectation given the observed data;  $\sum_{(i,k) \in R^c} E_{\text{re}}\{\partial \log\{1 - p_r(y_{ik}, x_{ik}; \gamma)\} / \partial \gamma \mid O, (i,k) \in R^c\}$ .

Orchard and Woodbury (1972) formalized 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  $\theta$ . If the missing data were actually observed, we could estimate  $\theta$  by the score function that is obtained from  $f$ . (The score (3.1) in the present case.) But, since  $M$  is unobserved, factorize  $f(O, M; \theta) = f_1(O; \theta) f_2(M \mid O; \theta)$  and estimate  $\theta$  from the marginal distribution  $f_1$  of the observed data, and the expectation  $E_2(M \mid O; \theta)$ .

Returning to the missing information principle equations (3.2), the vector parameter  $\gamma$  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

$$\begin{aligned}
& \sum_{(i,j) \in R} \log\{p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)\} \\
&+ \sum_{(i,k) \in R^c} E_s \left( \frac{E_{\text{re}}[\{p_r^{-1}(y_{ik}, \text{[redacted]}) - 1\} \log\{1 - p_r(y_{ik}, x_{ik}; \gamma)\} \mid x_{ik}, u_i, (i,k) \in R]}{E_{\text{re}}[\{p_r^{-1}(y_{ik}, x_{ik}; \gamma^*) - 1\} \mid x_{ik}, u_i, (i,k) \in R]} \middle| O \right). \tag{3.3}
\end{aligned}$$

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

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

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 equation (3.2).

*Remark 2.* When the response probabilities  $p_r(y_{ij}, \mathbf{x}_{ij}; \gamma)$  depend on only  $\mathbf{x}_{ij}$  (and  $\gamma$ ), they are referred to as *propensity scores*, and the missing data are MAR. This kind of response mechanism may hold in establishment survey settings, e.g. when the response propensity is related to the unit size. The estimating equations (3.2) reduce in this case to the common log-likelihood equations

$$\sum_{(i,j) \in R} \frac{\partial \log\{p_r(\mathbf{x}_{ij}; \gamma)\}}{\partial \gamma} + \sum_{(i,k) \in R^c} \frac{\partial \log\{1 - p_r(\text{ik} \gamma)\}}{\partial \gamma} = 0, \quad (3.5)$$

where  $p_r(\mathbf{x}_{ij}; \gamma) = \Pr(R_{ij} = 1 | \mathbf{x}_{ij}; \gamma)$ .

*Remark 3.* A fundamental question regarding the solution of the missing information principle equations (3.2) is the existence of a unique solution or, more generally, the identifiability of the response model. Recently, Riddles *et al.* (2016) proposed a similar approach to deal with NMAR non-response in the general context of sample surveys and established the following fundamental condition for the response model identifiability: the covariates  $\mathbf{x}$  can be decomposed as  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$  with  $\dim(\mathbf{x}_2) \geq 1$ , such that  $\Pr(R_{ij} = 1 | y_{ij}, \mathbf{x}_{ij}) = \Pr(R_{ij} = 1 | y_{ij}, \mathbf{x}_{1ij})$ . In other words, the covariates in  $\mathbf{x}_2$  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  $\mathbf{x}_2$  explain the variability of the outcome values and hence they provide valuable information on the missing values, and they 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, interestingly, 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  $\mathbf{x}_2$ . (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.

### 3.1. Example: mixed logistic model for outcome variable

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

$$p_y(x_{ij}, u_i) = \Pr\{y_{ij} = 1 | x_{ij}, u_i, (i, j) \in R; \beta\} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_i)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_i)}, \quad \text{ik} \quad u_i \stackrel{\text{IID}}{\sim} N(0, \sigma_u^2). \quad (3.6)$$

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

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


$$\begin{aligned} E_{re} \left[ \{p_r^{-1}(y_{ij}, x_{ij}; \gamma) - 1\} \frac{\partial \log\{1 - p_r(y_{ij}, x_{ij}; \gamma)\}}{\partial \gamma} \middle| x_{ij}, u_i, (i, j) \in R \right] \\ = p_y(x_{ij}, u_i) \{p_r^{-1}(1, x_{ij}; \gamma) - 1\} \frac{\partial \log\{1 - p_r(1, x_{ij}; \gamma)\}}{\partial \gamma} \\ + \{1 - p_y(x_{ij}, u_i)\} \{p_r^{-1}(0, x_{ij}; \gamma) - 1\} \frac{\partial \log\{1 - p_r(0, x_{ij}; \gamma)\}}{\partial \gamma}, \\ E_{re}[\{p_r^{-1}(y_{ij}, x_{ij}; \gamma) - 1\} | x_{ij}, u_i, (i, j) \in R] = p_y(x_{ij}, u_i) \{p_r^{-1}(1, x_{ij}; \gamma) - 1\} \\ + \{1 - p_y(x_{ij}, u_i)\} \{p_r^{-1}(0, x_{ij}; \gamma) - 1\}. \end{aligned}$$

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

*Remark 4.* A possible criticism of our proposed approach is that it requires the specification of 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 non-respondents. Although this is generally true, we mention that Rivers (2007) and Feder and Pfeiffermann (2015) defined 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 the use of 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 non-response.


#### 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  $\pi_i$  and  $\tilde{\pi}_{k|i} = \pi_{k|i} p_r(y_{ik}, \hat{y}_i)$ .

By Pfeiffermann and Sverchkov (2007), the optimal small  predictor for area  $i$  is

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

(It follows, from the identity,  $E_U\{(\hat{Y}_i - \bar{Y}_i)^2 | O, I_i\} = \{\hat{Y}_i - E(\bar{Y}_i | O, I_i)\}^2 + \text{var}_U(\bar{Y}_i | O, I_i)$ , for any predictor  $\hat{Y}_i$ .) We estimate therefore the area means in *sampled areas* as

$$\begin{aligned} \hat{Y}_i = \hat{E}_U(\bar{Y}_i | O, I_i = 1) = N_i^{-1} \left\{ \sum_{j: (i, j) \in R} y_{ij} + \sum_{k=1, k \notin R}^{N_i} \hat{E}_U(y_{ik} | O, I_i = 1) \right\} \\ = N_i^{-1} \left( \sum_{j: (i, j) \in R} y_{ij} + \sum_{k=1, k \notin R}^{N_i} E_s \left[ \frac{E_{re}\{(\tilde{\pi}_{k|i}^{-1} - 1) y_{ik} \text{  u_i, (i, k) \in R\}}{E_{re}\{(\tilde{\pi}_{k|i}^{-1} - 1) | x_{ik}, u_i, (i, k) \in R\}} \middle| O \right] \right) \\ = N_i^{-1} \left\{ \sum_{j: (i, j) \in R} y_{ij} + \sum_{k=1, k \notin R}^{N_i} E_s \left( \frac{E_{re}[\{\tilde{w}(y_{ik}, x_{ik}) - 1\} y_{ik} | x_{ik}, u_i, (i, k) \in R]}{E_{re}[\{\tilde{w}(y_{ik}, x_{ik}) - 1\} | x_{ik}, u_i, (i, k) \in R]} \middle| O \right) \right\}, \quad (4.2) \end{aligned}$$

where  $\tilde{w}(y_{ik}, \text{[icon]}) = E_{\text{re}}\{\tilde{\pi}_{k|i}^{-1} | y_{ik}, \text{[icon]}(i, k) \in R\}$ . (The second row follows from equation (2.3). We assume that  $E_{\text{re}}\{\tilde{\pi}_{k|i}^{-1} | y_{ik}, \text{[icon]}(i, k) \in R\} = E_{\text{re}}\{\tilde{\pi}_{k|i}^{-1} | y_{ik}, \text{[icon]}(i, k) \in R\}$ .) The external expectation in the last row of equation (4.2) is over the distribution of  $u_i$  under the sample model (no non-response 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 equation (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  $\pi_i$  and  $\tilde{\pi}_{k|i} = \pi_{k|i} \hat{p}_r(y_{ik}, \text{[icon]}) = \pi_{k|i} p_r(y_{ik}, \text{[icon]})$ .

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

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

The prominent feature of this estimator is that it uses the estimated probabilities  $\tilde{\pi}_{j|i} = \pi_{j|i} p_r(y_{ij}, \mathbf{x}_{ij}; \hat{\gamma})$ . 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

$$\hat{Y}_i = E_U(\bar{Y}_i | O, I_i = 0) = N_i^{-1} \left\{ \sum_{k=1}^{N_i} E_U(y_{ik} | O, I_i = 0) \right\} = N_i^{-1} \sum_{k=1}^{N_i} \frac{\sum_{l \in s} \{(\pi_l^{-1} - 1) K_l(\text{[icon]})\}}{\sum_{l \in s} (\pi_l^{-1} - 1)}, \quad (4.4)$$

where

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

See Pfeffermann and Sverchkov (2007), section 7, for a derivation of equation (4.4).

*Remark 6.* Recently, Verret *et al.* (2015) proposed to account for informative sampling within the areas by including the sampling weights or functions of them as additional explanatory variables in the model. (They assumed that all the areas are sampled with full response.) However, a similar approach cannot be used to account for NMAR non-response 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  $\hat{p}_r(y_{ik}, \text{[icon]})$  can be computed only for the responding units.

## 5. Mean-squared error estimation

We propose a semiparametric bootstrap procedure for MSE estimation. The procedure uses the same idea as in Sverchkov and Pfeffermann (2004). We first generate a pseudopopulation with



marginal distributions of the outcome values, similar to the distributions of the true population values, and then select independently  $B$  samples from the pseudopopulation by 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 on the basis of the sample of respondents.

### 5.1. Generation of pseudopopulation

- (a) Use the observed data 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 that are 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  $\tilde{\pi}_i = 1/\tilde{w}_i$ .
- (b) Generate a synthetic population with values  $\tilde{y}_{ij} = \hat{p}_y(\mathbf{x}_{ij}, \tilde{u}_i)$  and  $\tilde{\pi}_{j|i} = 1/\hat{w}(\tilde{y}_{ij}, \mathbf{x}_{ij})$ ; ( $\hat{w}(\tilde{y}_{ik}, \mathbf{x}_{ik})$  is computed as below equation (4.2) with  $\tilde{y}_{ik}$  instead of  $y_{ik}$ ) and response probabilities  $\tilde{p}_{ij}^{\text{resp}} = p_r(\tilde{y}_{ij}, \mathbf{x}_{ij}; \hat{\gamma})$ ,  $i = 1, \dots, M$ ,  $j = 1, \dots, N_i$ . 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.
- (c) For each area  $i = 1, \dots, M$  of size  $N_i$  in the synthetic population, sample *with replacement*  $N_i$  units with probabilities proportional to  $1/(\tilde{\pi}_{j|i})$ .

This concludes the generation of the pseudopopulation.

*Remark 7.* As implied by the results of Sverchkov and Pfeiffermann (2004), if the model hyperparameters and random effects were actually known, the marginal distributions of the outcomes  $\tilde{y}_{ij}$  in the pseudopopulation would have been the same as the corresponding marginal distributions of the outcomes  $y_{ij}$  in the original population. In practice, one can use only estimated parameters but, as our simulation study shows, the procedure that is proposed in this section for MSE estimation, which relies on generating the pseudopopulation, performs well even for areas with small samples. Note in this respect that the model hyperparameters are estimated from all the sampled areas but, for given hyperparameter estimates, the estimates of the random effects are ‘direct’.

### 5.2. Selection of bootstrap samples and computation of estimates

- (a) Sample independently  $B$  samples  $(\tilde{y}_{ij}^b, \mathbf{x}_{ij}^b, \tilde{\pi}_{j|i}^b, \tilde{\pi}_i^b)$ ,  $b = 1, \dots, B$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , from the pseudopopulation by using the same sampling schemes as used for selecting the original sample, but with inclusion probabilities  $\tilde{\pi}_i$  and  $\tilde{\pi}_{j|i}$ .
- (b) For each unit in the sample define the response probability as  $p_r(\tilde{y}_{ij}^b, \mathbf{x}_{ij}^b; \hat{\gamma})$ , where  $\hat{\gamma}$  is the estimate that is obtained from the true original sample.
- (c) For each bootstrap sample  $b$ , re-estimate all the parameters of interest (means or totals in the present paper).
- (d) Calculate empirical MSE or other statistics of interest over the  $B$  bootstrap samples.

As implied by the description of the bootstrap method proposed, 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 that were 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 six steps.

*Step 1* (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

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

$i = 1, \dots, 300$ ,  $N_i = \text{int}(1000 \exp[\min\{2.5, \max(-2.5, u_i^U)\}/5])$ . 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  $\sigma_u^2 = 0.01$ , to study the effect of the magnitude of the variance on the performance of alternative estimators (see below).

Group the areas randomly into three sets:  $G1 = \{i = 1, \dots, 100\}$ ;  $G2 = \{i = 101, \dots, 200\}$ ;  $G3 = \{i = 201, \dots, 300\}$ .

*Step 2* (sample selection): select 50 areas from each group by systematic probability proportional to size sampling with the area sizes  $N_i$  as the size variable. Note 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 group G1, 40 units from each selected area in group G2 and 60 units from each selected area in group G3 by using systematic probability proportional to size 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  $y_{ij}$ .

*Step 3* (response mechanism): obtain response from unit  $j$  in sampled area  $i$  with probability

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

where  $\gamma_0 = 0$ ,  $\gamma_1 = -0.5$  and  $\gamma_2 = 2$ . The non-response 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 (equation (6.8) below). For this case the response rate was only 46%.

*Step 4* (fitting of respondents' 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

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

using PROC NLMIX in SAS with default options. Note that 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).

*Step 5* (estimation of response probabilities): assume that  $p_r(y_{ij}, x_{ij}, \gamma) = \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij}) / \{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})\}$ , compute the expectations in equation (3.2) under the estimated model  $\hat{p}_y(x_{ij}, \hat{u}_i)$  in model (6.3) and solve the resulting equations to estimate  $\gamma$ , using the procedure that was described in Section 3.

*Step 6* (prediction of area means): first estimate  $\tilde{w}(y_{ij}, x_{ij}) = E_{\text{re}}\{\tilde{\pi}_{j|i}^{-1} | y_{ij}, x_{ij}, (i, j) \in R\}$  as follows. By definition,  $E_{\text{re}}\{\tilde{\pi}_{j|i}^{-1} | y_{ij}, x_{ij}, (i, j) \in R\} = p_r^{-1}(y_{ij}, x_{ij}) E_{\text{re}}\{\pi_{j|i}^{-1} | y_{ij}, x_{ij}, (i, j) \in R\}$ , where  $\pi_{j|i} = n_i z_{ij} / \sum_{j=1}^{N_i} z_{ij} = (n_i / N_i) z_{ij} (1 / \bar{Z}_i)$  and  $\bar{Z}_i = (1 / N_i) \sum_{j=1}^{N_i} z_{ij}$  is the  $i$ th area mean, which is viewed as a constant, assuming that the true area size is large. Let  $z_{ij}^* = \pi_{j|i} N_i / n_i$ . In the present simulation study we fit the model  $z_{ij}^* = g_\alpha(y_{ij}, x_{ij}) = \alpha_0 + \alpha_y y_{ij} + \alpha_x x_{ij} + \varepsilon_{ij}$ , but other models can be fitted, depending on the available data. Note that  $g_\alpha(y_{ij}, x_{ij})$  refers to the sample data before response and therefore the response weights  $p_r^{-1}(y_{ij}, x_{ij}; \hat{\gamma})$  must be used for estimating this model via weighted regression. Alternatively, we 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  $\tilde{w}(y_{ij}, x_{ij})$  as

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

Next, compute the ratios of the estimated expectations

$$\widetilde{\text{Ra}}(x_{ik}, u_i) = \frac{E_{\text{re}}[\{\hat{\tilde{w}}(y_{ik}, x_{ik}) - 1\} y_{ik} | x_{ik}, u_i, (i, k) \in R]}{E_{\text{re}}[\{\hat{\tilde{w}}(y_{ik}, x_{ik}) - 1\} | x_{ik}, u_i, (i, k) \in R]} \quad (6.5)$$

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 equation (6.5) into equation (4.2) 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.* For brevity, we consider only the prediction of the mean outcome in sampled areas, which are subject to NMAR non-response. The estimation of the means of non-sampled areas is the same as in Pfeiffermann and Sverchkov (2007) and is illustrated in their simulation study.

*Repeat steps 1–6 independently 500 times.* The values  $x_{ij}$  of the covariate are generated only once and held fixed for all the simulations.

(a) Predictors considered: compute the following predictors for each simulation.

- (i)  $\hat{Y}_i^{\text{ign}} = N_i^{-1} \{\sum_{j, (i, j) \in R} y_{ij} + \sum_{k=1, k \notin R}^{N_i} \hat{p}_y(x_{ij}, u_i)\}$  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.
- (ii)  $\hat{Y}_i^{\text{HB, MCAR}} = \sum_{j, (i, j) \in R} \pi_{j|i}^{-1} y_{ij} / \sum_{j, (i, j) \in R} \pi_{j|i}^{-1}$ ; this is the familiar Hajek–Brewer (Hajek, 1971) estimator that ‘assumes’ that the non-response is completely at random.
- (iii)  $\hat{Y}_i^{\text{MAR}} = \sum_{j, (i, j) \in R} \hat{w}(x_{ij}) y_{ij} / \sum_{j, (i, j) \in R} \hat{w}(x_{ij})$ ;  $\hat{w}(x_{ij}) = \{\pi_{j|i} p(x_{ij}, \hat{\lambda})\}^{-1}$ ; this estimator accounts for the response process but assumes that the non-response is MAR, and hence the response probabilities are estimated by assuming the propensity scores model  $\Pr(R_{ij} = 1 | x_{ij}; \lambda) = p_r(x_{ij}, \lambda) = \exp(\lambda_0 + \lambda_1 x_{ij}) / \{1 + \exp(\lambda_0 + \lambda_1 x_{ij})\}$ . The parameter  $\lambda = (\lambda_0, \lambda_1)'$  is estimated by solving the likelihood equations

$$\sum_{(i, j) \in R} \frac{\partial \log\{p_r(x_{ij}; \lambda)\}}{\partial \lambda} + \sum_{(i, j) \in R^c} \frac{\partial \log\{1 - p_r(x_{ij}; \lambda)\}}{\partial \lambda} = 0.$$

- (iv)  $\hat{Y}_i^{\text{HB}} = \sum_{j, (i, j) \in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j, (i, j) \in R} (1 / \tilde{\pi}_{j|i})$ ; this has already been defined in equation (4.3) and accounts for NMAR non-response.

- (v)  $\hat{Y}_i^{\text{new}} = N_i^{-1}(\sum_{j,(i,j) \in R} y_{ij} + \sum_{k=1, k \neq i}^{N_i} \hat{y}_{ik})$ ; this is the proposed empirical model-dependent predictor obtained from Equation (4.2). The ratios  $\hat{y}_{ik}$  are obtained from model (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 non-response.

- (b) The statistics that are considered for assessment of performance of predictors and RMSE estimates are as follows.
- (i) *Prediction of area means*: let  $D_{ir} = 1$  or  $D_{ir} = 0$  if area  $i$  is respectively sampled or 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  $\hat{Y}_{ir}$  represent any of the five predictors defined above,  $r = 1, \dots, 500$ :

$$\text{Bias}_i = \frac{\sum_{r=1}^{500} D_{ir} (\hat{Y}_{ir} - \bar{Y}_{ir})}{\sum_{r=1}^{500} D_{ir}},$$

$$\text{RMSE}_i = \sqrt{\left\{ \frac{\sum_{r=1}^{500} D_{ir} (\hat{Y}_{ir} - \bar{Y}_{ir})^2}{\sum_{r=1}^{500} D_{ir}} \right\}}.$$
(6.6)

- (ii) *Estimation of the RMSE*: because of running time limitations, for estimation of the RMSE we considered only the first 100 simulations and generated only 50 bootstrap samples for each simulation. Let  $D_{irb} = 1$  or  $D_{irb} = 0$  if area  $i$  is respectively sampled or not sampled for 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  $\hat{Y}_{irb}^{\text{new}}$  represent the corresponding new predictor:

$$\text{RMSE}_i^{\text{Boot}} = \sqrt{\left( \frac{1}{100} \sum_{r=1}^{100} \widehat{\text{MSE}}_{i,r}^{\text{Boot}} \right)},$$

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

In any given application, one would obviously generate many more bootstrap samples but we report summary statistics over the 100 simulations, so we actually report the results obtained over  $100B_i$  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.

As implied by the results in Table 1, although the estimators of all three coefficients are biased, the biases are relatively very small and so are the standard deviations  $S_{\hat{\theta}_k}$  of the estimators. 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

**Table 1.** Estimation of response model coefficients

<i>Coefficient</i>	<i>Bias</i>	<i>Std</i>
$\gamma_0 = 0$	0.006	0.055
$\gamma_1 = -0.5$	0.003	0.045
$\gamma_2 = 2$	0.037	0.174


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 fifth area.

The conclusions from Fig. 1 are clear cut. The proposed model-dependent predictor  $\hat{Y}_i^{\text{new}}$  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  $\hat{Y}_i^{\text{ign}}$  that ignores both the informative sampling and the response.

The RMSE of our proposed predictor,  $\hat{Y}_i^{\text{new}}$ , is uniformly the smallest, with the Hajek–Brewer estimator being second in order (Fig. 2). The RMSE of  $\hat{Y}_i^{\text{ign}}$  is dominated by its large bias and hence its large value. The RMSEs of all the predictors decrease as the sample sizes increase, because of the decrease in the variance.

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

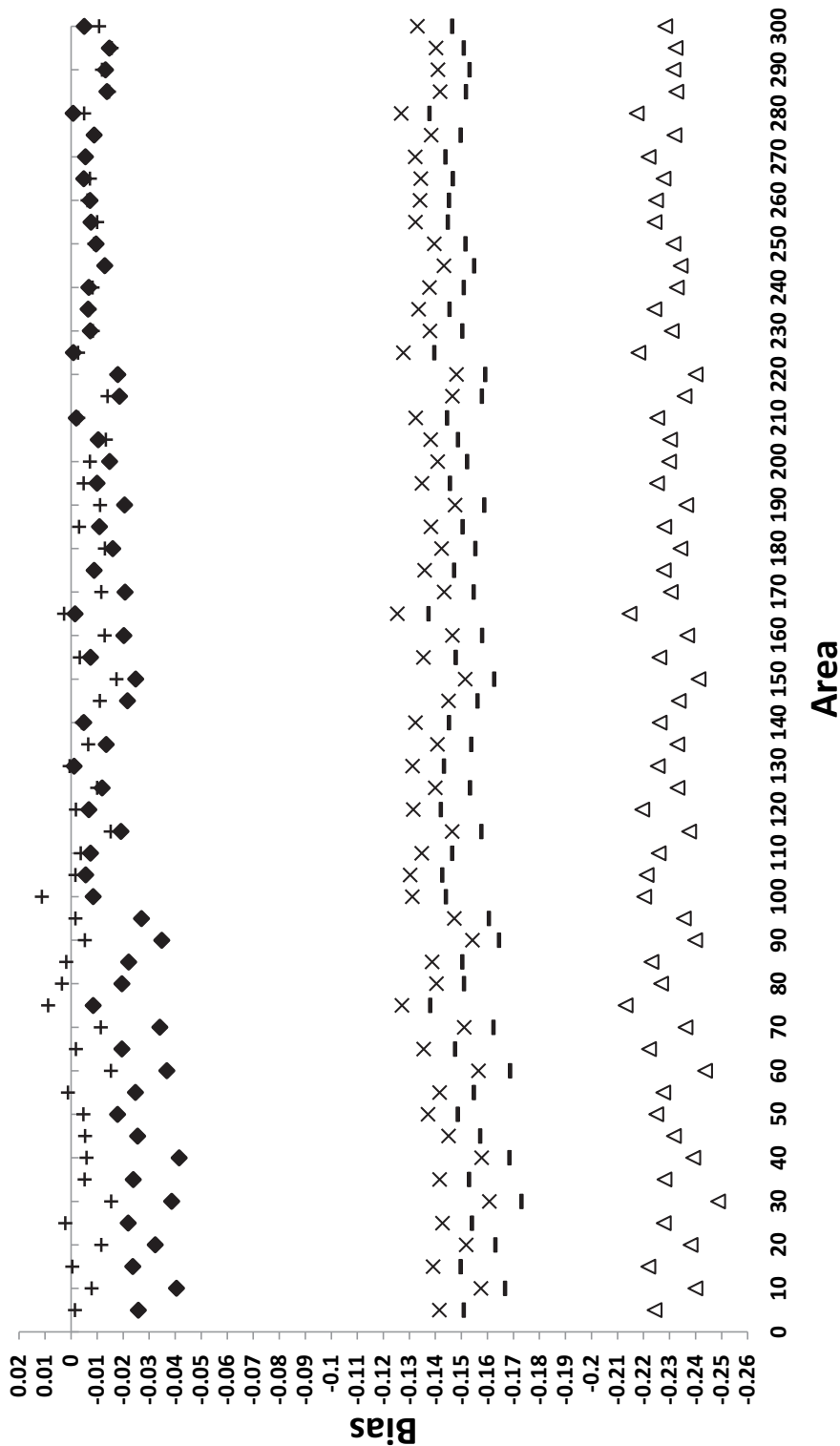
 Next we illustrate the robustness of the model that is assumed for the response probabilities, discussed in remark 4. For this, we repeated the same simulation study but with a different true response model:

$$p(y_{ij}, x_{ij}, \gamma) = \frac{\exp(-0.5x_{ij}y_{ij})}{1 + \exp(-0.5x_{ij}y_{ij})}. \quad (6.8)$$

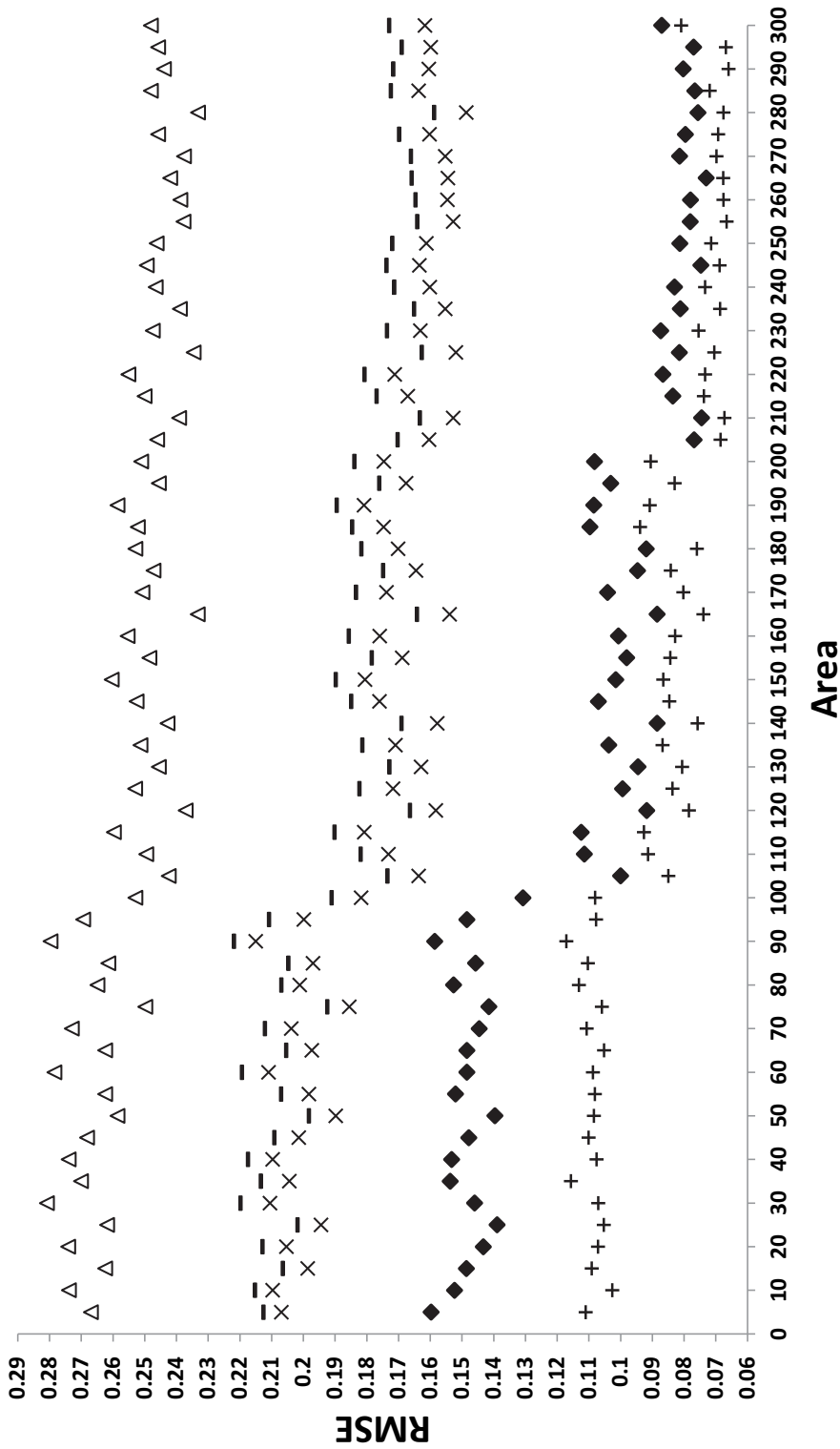
(Compare with equation (5.2).) However, we fit the response model  $p_r(y_{ij}, x_{ij}, \gamma) = \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij}) / \{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})\}$ . (This is the same as before. We did not add the cross-product  $x_{ij}y_{ij}$  to the working model because it would make the true response model a special case and we want to illustrate the robustness of the working response model. Note also that  $X$  and  $Y$  are binary, so there is no point in 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.

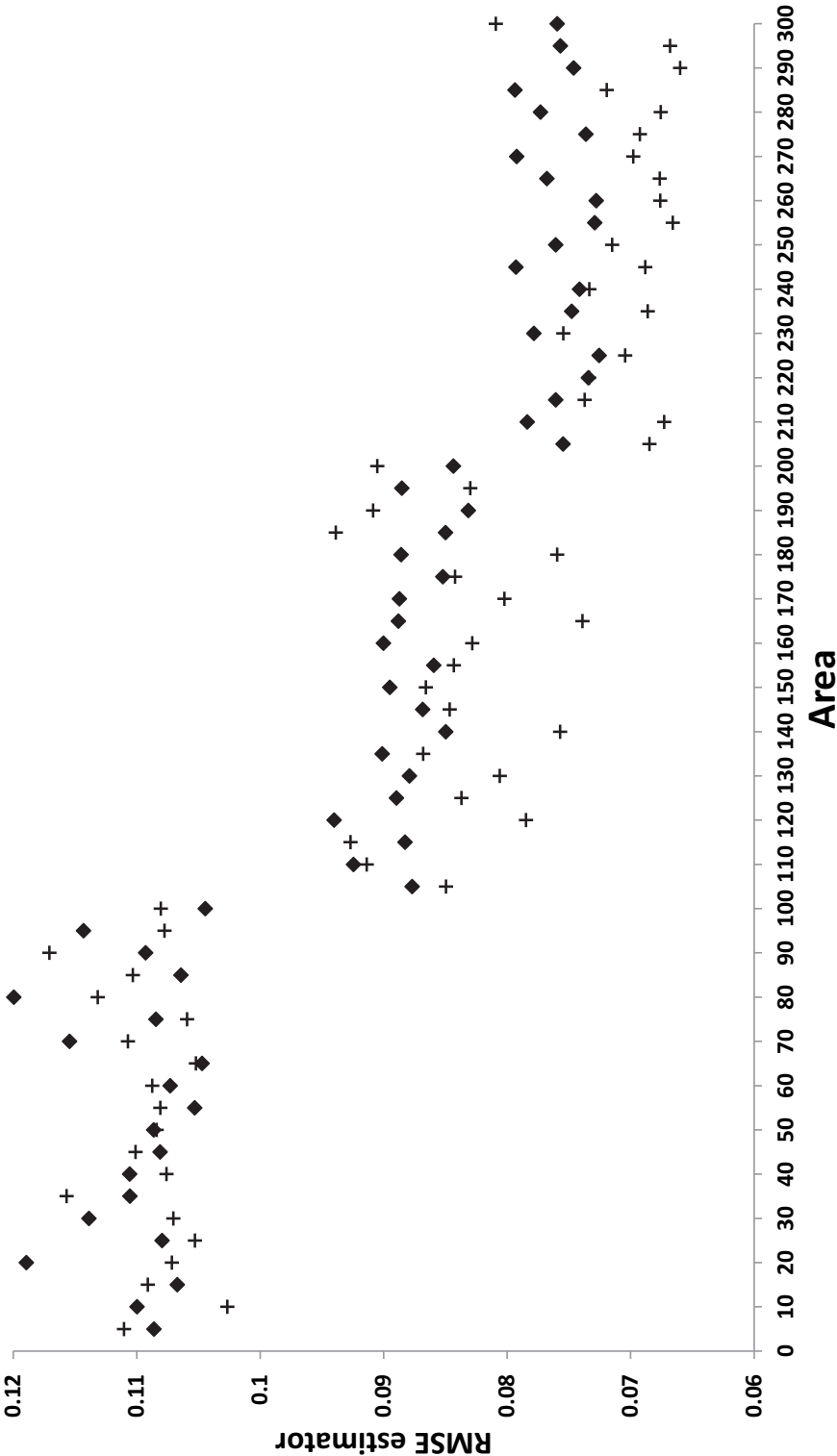
Fig. 4 exhibits a similar picture to Fig. 1, with the new and the Hajek–Brewer predictors now being slightly biased. The other three predictors are more biased, but the bias is considerably smaller than in Fig. 1, as obtained when estimating the correct response model. The different



**Fig. 1.** Bias of predictors by area, 500 simulations ( $\sigma_{ij}^2 = 1$ ):  $\Delta$ , ignorable; —, MAR;  $\times$ , Hajek-Brewer, missing completely at random;  $\blacklozenge$ , Hajek-Brewer;  $+$ , new

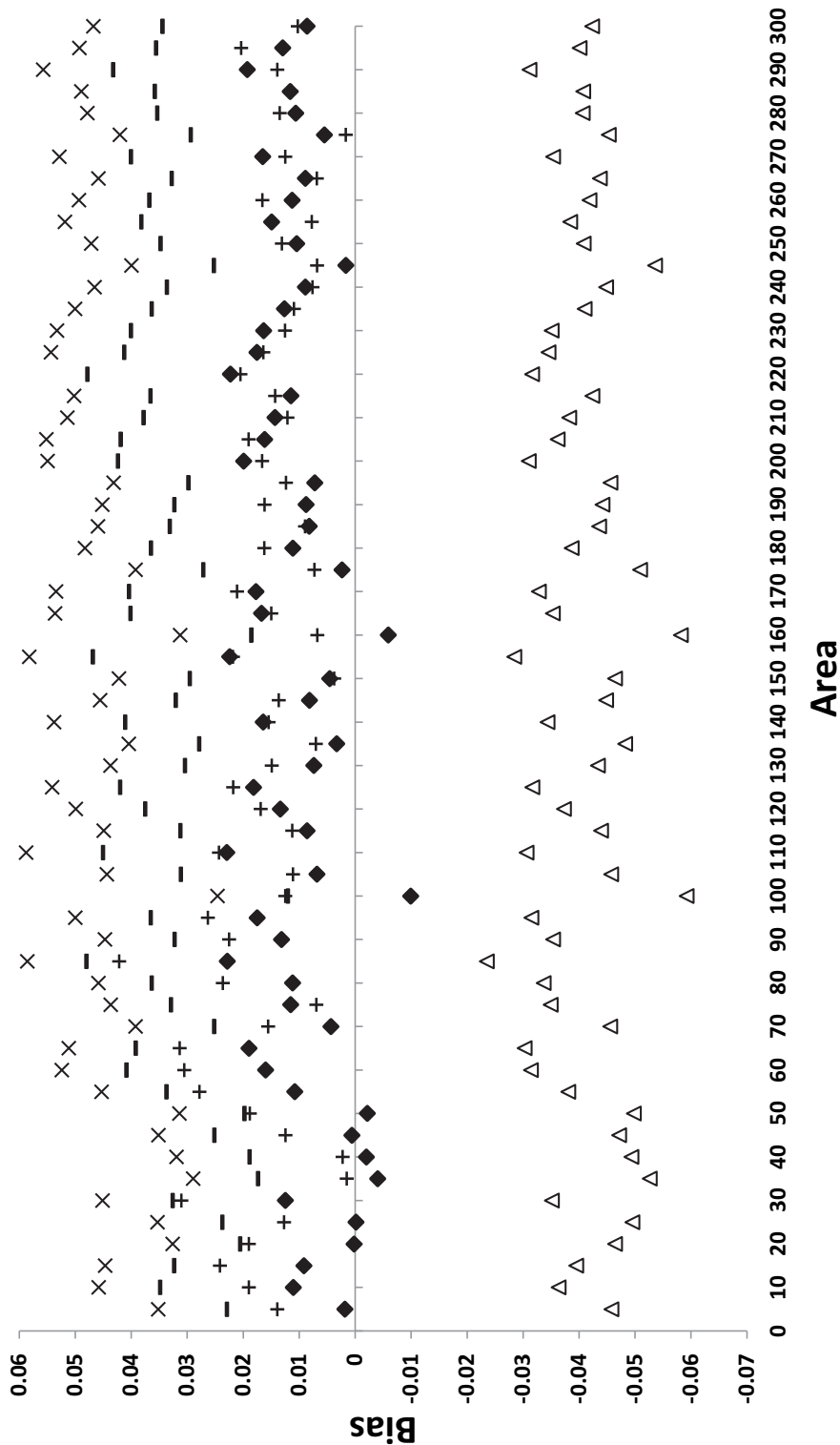


**Fig. 2.** RMSE of predictors by area, 500 simulations ( $\sigma_u^2 = 1$ ):  $\Delta$ , random;  $\times$ , Hajek-Brewer, missing completely at random;  $\diamond$ , Hajek-Brewer, new;  $+$ , new



**Fig. 3.** Estimation of RMSE of  $\hat{Y}_i^{\text{new}}$  by area ( $\sigma_{ij}^2 = 1$ ): +, empirical RMSE over 500 simulations; ◆, mean over the first 100 simulations, with 50 bootstrap samples for each simulation





**Fig. 4.** Bias of predictors by area, response model misspecified, 500 simulations:  $\Delta$ , ignorables; —, MAR; X, Hajek-Brewer, missing completely at random;  $\diamond$ , Hajek-Brewer; + new

magnitudes of the bias of the three predictors in Figs 1 and 4 are 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 non-response. Thus, Figs 1 and 4 are not really comparable.

The RMSEs of the proposed and the Hajek–Brewer predictors change only slightly when misspecifying the model of the response probabilities (Fig. 5). The RMSEs of the other three predictors are smaller under the misspecified model, because of the decrease in the bias.

Fig. 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.

As expected, the biases of all the three estimators are now larger, and so are the standard deviations (Table 2), 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 with 0.012 when  $\sigma_u^2 = 1$ ). Note 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 an 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 an average standard deviation of 0.221. The corresponding figures for the within-area sample selection probabilities are 0.0196 and 0.00572.

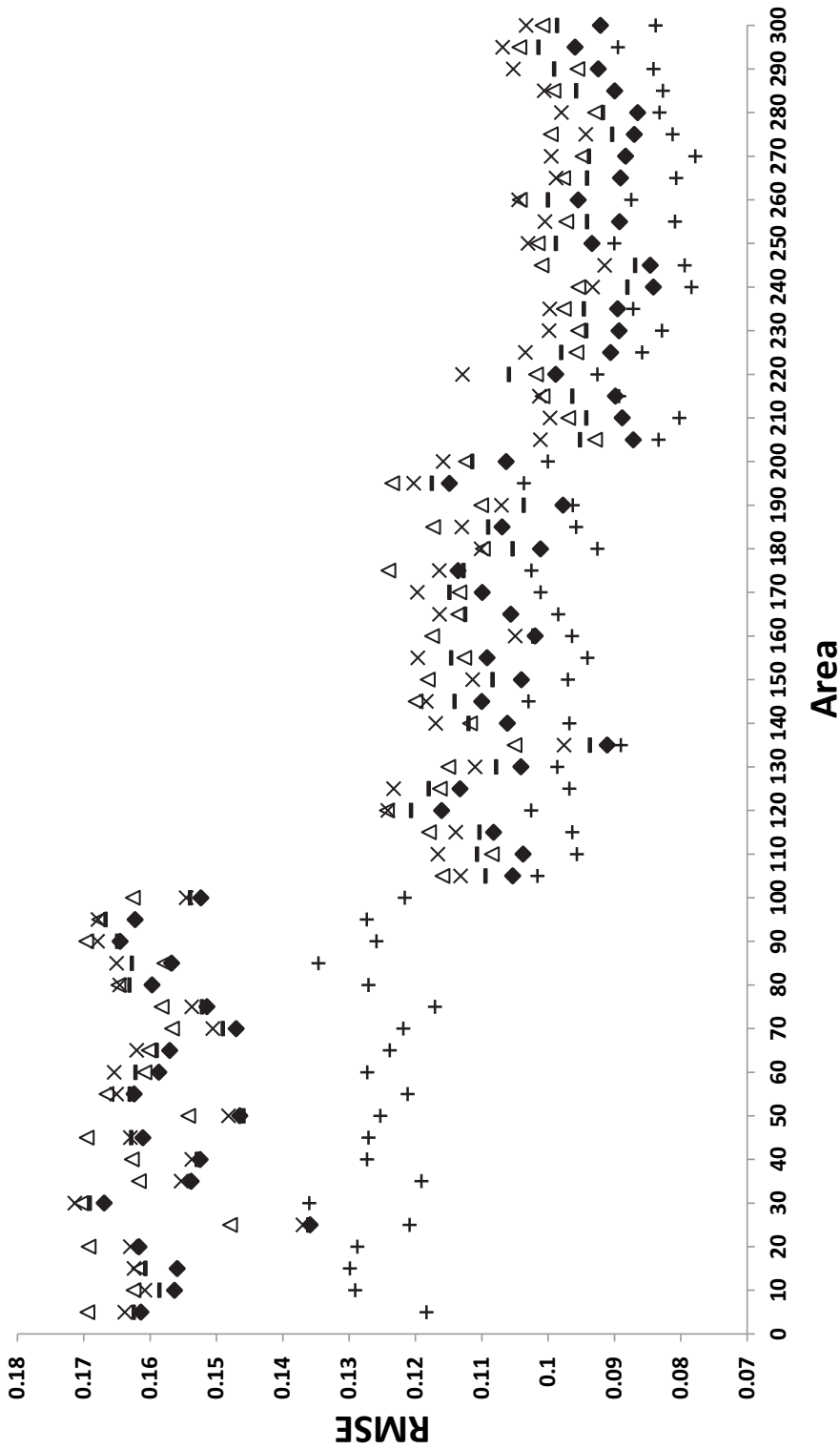
As in Section 6.2, the figures that follow illustrate the performance of the various predictors at the area level.

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

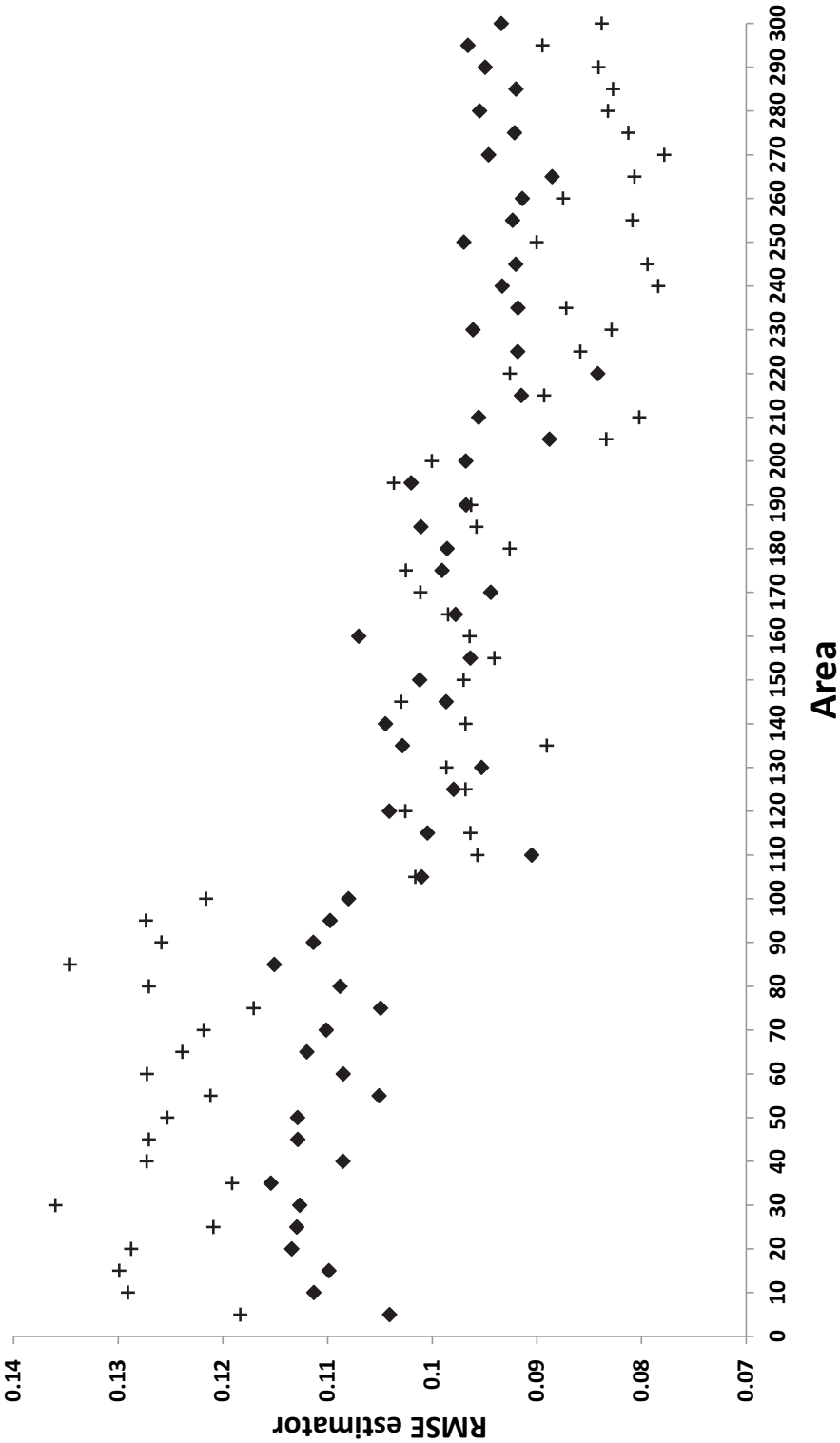
*Remark 9.* We repeated the simulation study also for the case  $\sigma_u^2 = 0.1$  and the two predictors  $\hat{Y}_i^{\text{HB}}$  and  $\hat{Y}_i^{\text{new}}$  still perform much better than the other three predictors that were considered,

**Table 2.** Estimation of response model coefficients

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



**Fig. 5.** RMSE of predictors by area, response model misspecified, 500 simulations:  $\Delta$ , ignorable;  $-$ , MAR;  $\times$ , Hajek-Brewer, missing completely at random;  $\diamond$ , Hajek-Brewer;  $+$ , new



**Fig. 6.** Estimation of  $RMSE(Y_{i'}^{new})$  by area, response model misspecified: +, empirical RMSE over 500 simulations; ◆, mean over the first 100 simulations, with 50 bootstrap samples for each simulation

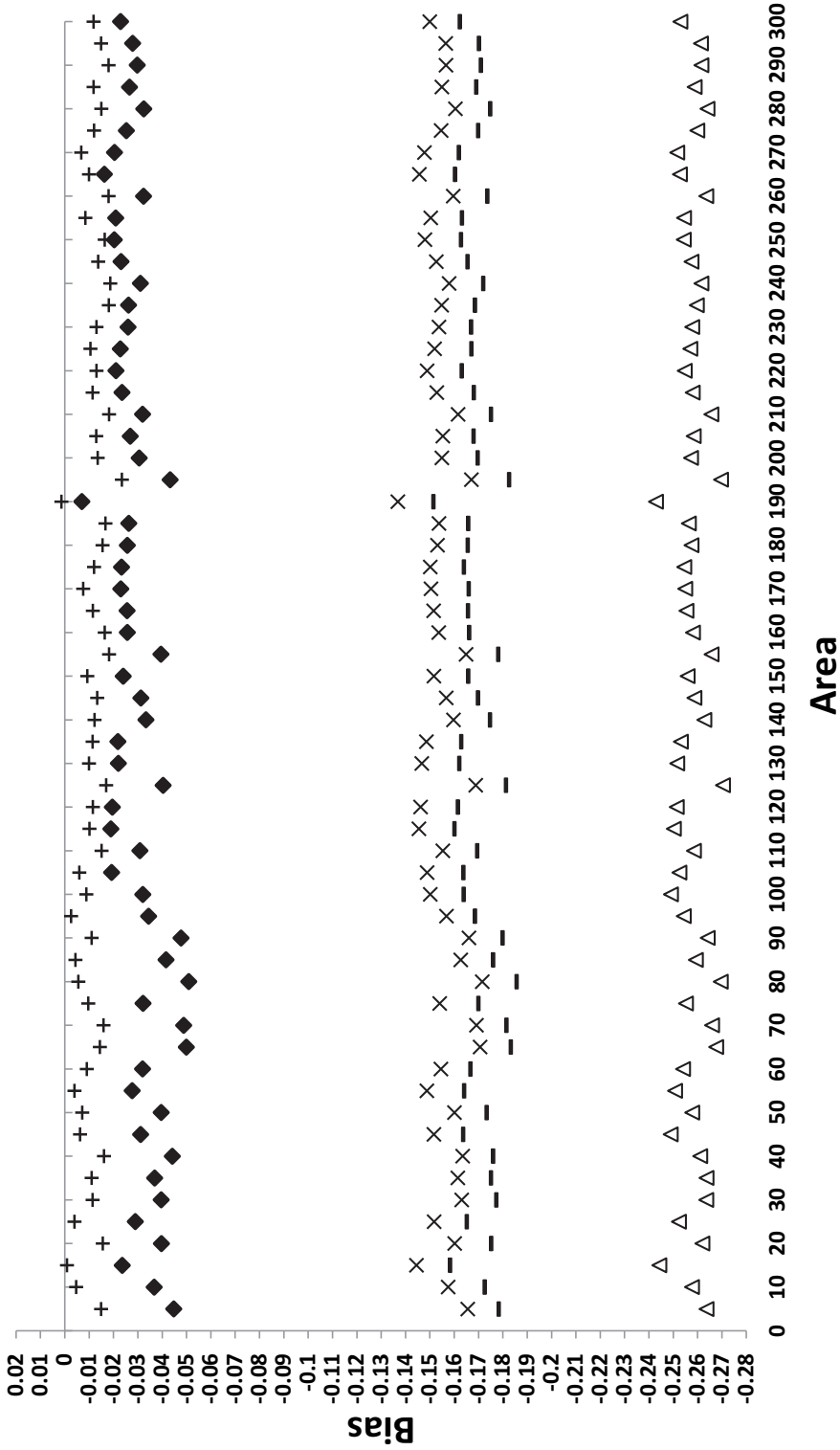


Fig. 7. Bias of predictors by area, 500 simulations ( $\sigma_{\theta}^2 = 0.25$ ):  $\Delta$ , ignorable;  $+$ , MAR;  $\times$ , Hajek-Brewer, missing completely at random;  $\diamond$ , Hajek-Brewer;  $+$ , new

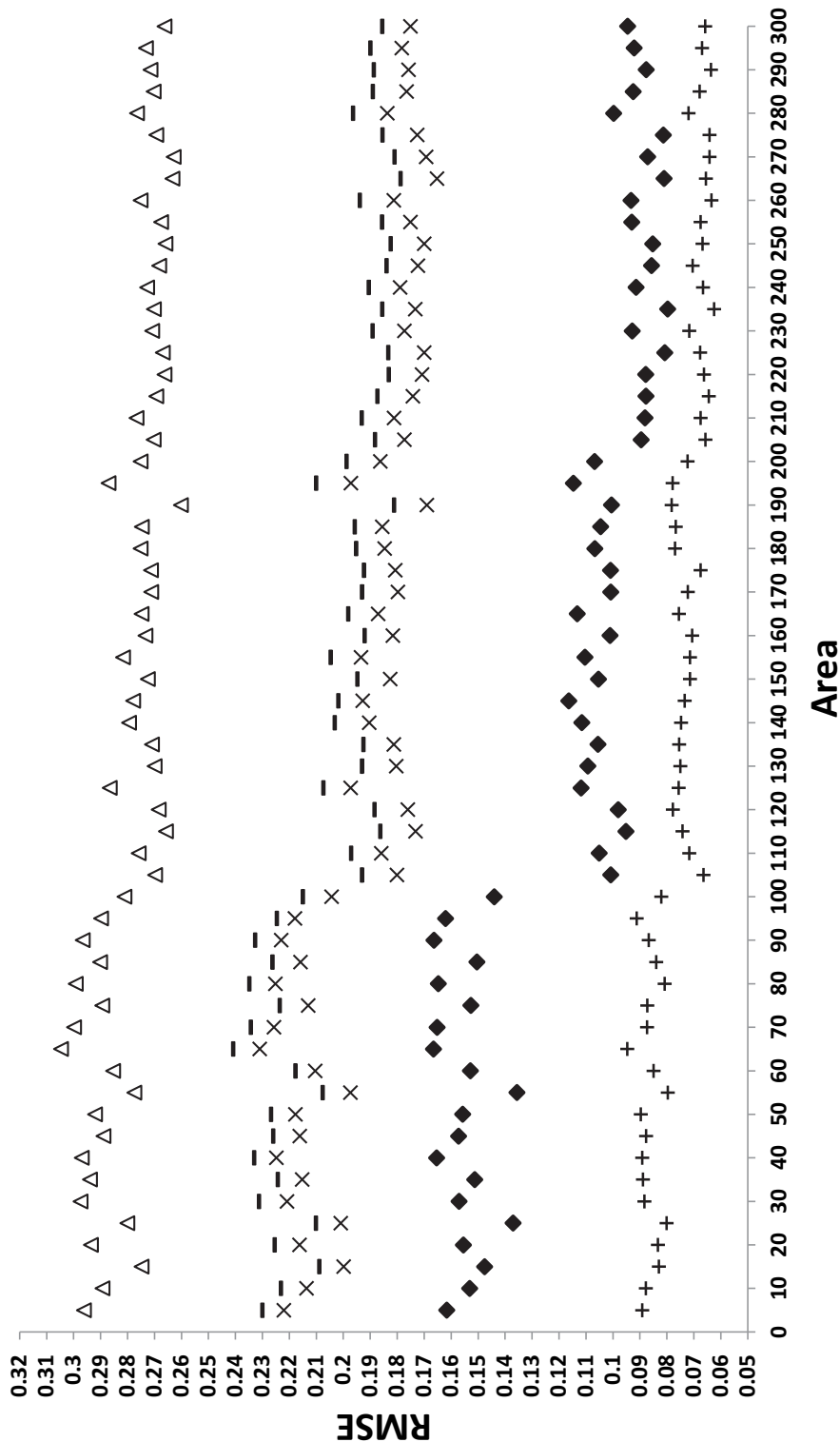
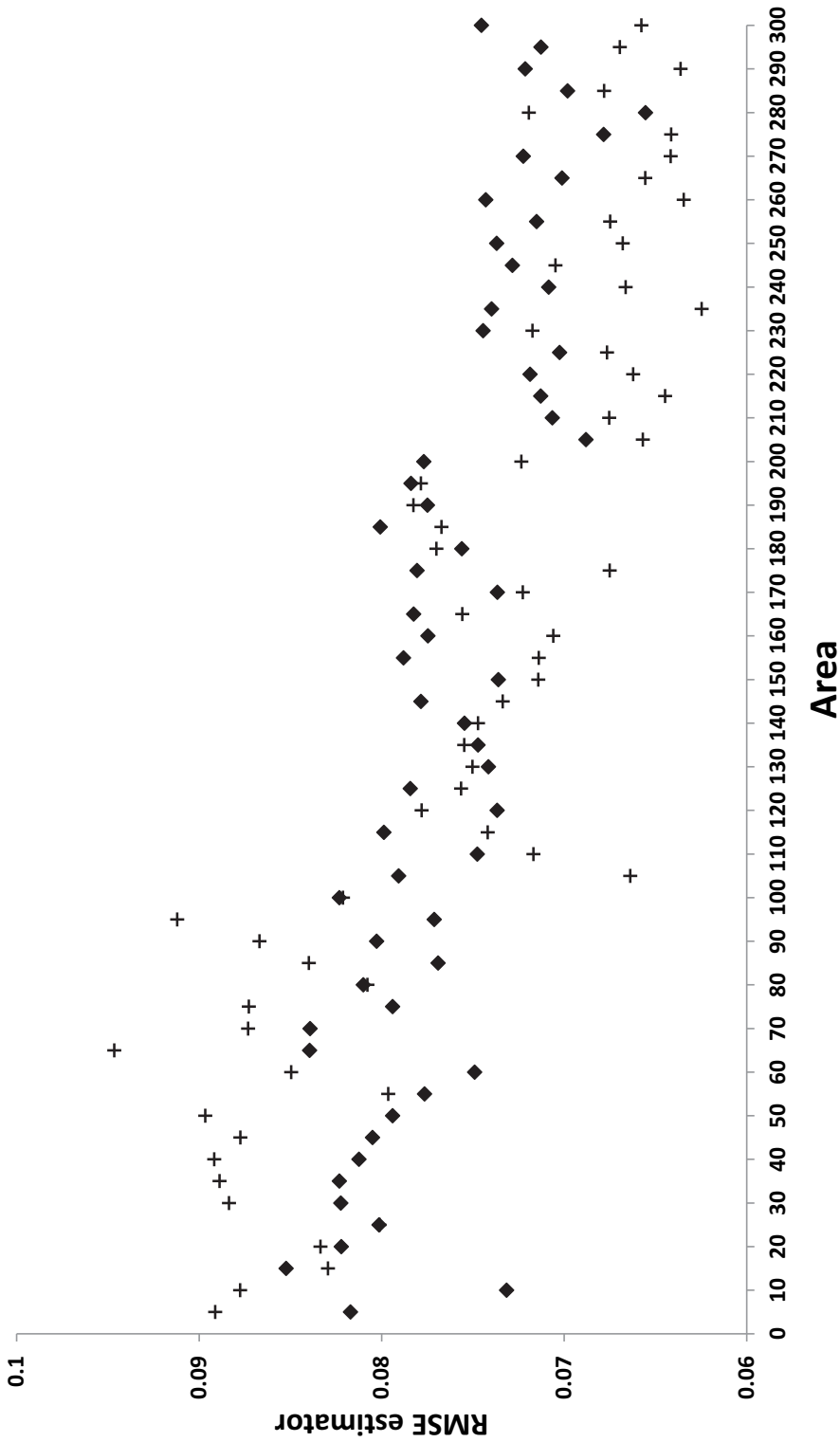


Fig. 8. RMSE of predictors by area, 500 simulations ( $\sigma_D^2 = 0.25$ ):  $\Delta$ , ignoreable;  $—$ , MAR;  $X$ , Hajek-Brewer, missing completely at random;  $\blacklozenge$ , Hajek-Brewer;  $+$ , new



**Fig. 9.** Estimation of the RMSE of  $\hat{Y}_i^{\text{new}}$  by area ( $\sigma_u^2 = 0.25$ ): +, empirical RMSE over 500 simulations; ♦, mean over the first 100 simulations, with 50 bootstrap samples for each simulation

although they now have a somewhat larger bias and RMSE. However, this is no longer so when  $\sigma_u^2 = 0.01$ , in which case all the five predictors perform badly because of failure to estimate the response probabilities properly, 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, to reduce the unexplained variations that are represented by the random effects (small  $\sigma_u^2$ ). However, if no other covariates  $\mathbf{x}_2$  that explain the outcome variable but not the response exist, then it is important that the variance  $\sigma_u^2$  of the random effects is not too small, thus enabling us to estimate the response probabilities and to remove the bias that is induced by NMAR non-response.

## 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, which is shared by other countries, is the enumeration of emigrants, as it is difficult to define emigrants and to count them. However, population counts are required for small domains, as defined by ‘statistical areas’, with an average size of about 3000 people. For these small domains, the population register is much less accurate, with an average enumeration error of about 13% and a 95th percentile of 40%. The main reason for the inaccuracy of the register at the statistical area level is that people moving into or out of 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 changes 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 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 who were registered in the area for estimating the *register overcount* (people registered falsely as living in the area). Denote  $N_i$  the true number of people living in area  $i$ ,  $K_i$  the number of people registered as living in area  $i$ ,  $p_{i, L|R}$  the proportion of people living in area  $i$  among those registered in the area and  $p_{i, R|L}$  the proportion of people registered to area  $i$  among those living in the area. Then,

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

Thus,  $N_i$  is estimated from the two samples as

$$\hat{N}_i = K_i \frac{\hat{p}_{i, L|R}}{\hat{p}_{i, R|L}}, \quad (7.2)$$

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

$$\text{var}(\hat{N}_i | K_i) = K_i^2 \left\{ \frac{\text{var}(\hat{p}_{i, L|R})}{E(\hat{p}_{i, L|R})^2} + \frac{E(\hat{p}_{i, L|R})^2}{E(\hat{p}_{i, R|L})^2} \text{var}(\hat{p}_{i, R|L}) \right\}. \quad (7.3)$$

In what follows we restrict ourselves to the overcount survey. Before the phone calls, a letter was sent to all the members sampled notifying them of the survey and asking them to respond to the phone interview. Nonetheless, there is a high rate of non-response in this survey, with




an average response rate of about 0.75 and a standard deviation between areas of about 0.14. Moreover, it is quite obvious that the non-response is NMAR because the non-respondents are more likely to be the people who are not registered correctly (living in another area) and hence who did not receive the notice letter in the first place.

The sampling design that was used in each statistical area is systematic sampling after ordering the frame by age. Note that this sampling scheme is *non-informative* since all the sampling units in a given area have the same sampling probability. The target is to estimate the total number of people who are 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 naive estimates, which ignore the non-response altogether. However, we have no information on the true target numbers of people who were registered correctly, so analysing 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 who were 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  $x_{ij} > 25$  and  $X_{1ij} = 0$  otherwise, and  $X_{2ij} = 1$  if  $x_{ij} > 40$  and  $X_{2ij} = 0$  otherwise. The following logistic models have been assumed for the outcomes that were observed for the responding units, and for the response probabilities, after removing from the data set people aged 16 years or less, who are all single.

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

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

Denote by  $R_i$  the subsample of responding people in the sample of people 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  people aged 16 years or older who were registered as living in the area:

$$\hat{M}_{i, \text{MCR}} = \frac{\sum_{j \in R_i} y_{ij}}{\sum_{j \in R_i} 1},$$

which assumes missing completely at random non-response;

$$\hat{M}_{i, \text{MR}} = \sum_{j \in R_i} y_{ij} + \sum_{k=1, k \notin R_i}^{\tilde{K}_i} \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1ik} + \hat{\beta}_2 X_{2ik} + \hat{u}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1ik} + \hat{\beta}_2 X_{2ik} + \hat{u}_i)},$$

which assumes MAR non-response;

$$\hat{M}_i^{\text{HB}} = \frac{\sum_{j \in R_i} y_{ij} \hat{p}_r^{-1}(y_{ij}, \mathbf{x}_{ij}; \gamma)}{\sum_{j \in R_i} \hat{p}_r^{-1}(y_{ij}, \mathbf{x}_{ij}; \gamma)},$$

which is the Hajek–Brewer estimator with estimated probabilities;

$$\hat{M}_i^{\text{new}} = \sum_{j \in R_i} y_{ij} + \sum_{k=1, k \notin R_i}^{\tilde{K}_i} \hat{\text{Ra}}_{ik},$$

which is the predictor proposed. See Section 6.1.

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

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

## 7.2. Results

We consider separately areas of size  $K_i \leq 100$  ( $A = 565$  areas, with an average size of 75 people) and areas of size  $100 < K_i \leq 200$  ( $A = 570$  areas, with an average size of 132 people), the small population areas. The mean number of responding units in the first group of areas is 63.48, with a 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 four predictors. Table 3 contains the following summary statistics for the four predictors over all the areas in each of the two groups:

$$Bias_{Err} = \sum_{i=1}^A E_i / A, \quad (7.6)$$

$$RMSE_{Err} = \left( \sum_{i=1}^A E_i^2 / A \right)^{0.5};$$

- (a) small areas ( $K_i \leq 100$ )— $\hat{\gamma}_0 = 0.56$ ,  $\hat{\gamma}_y = 0.67$ ,  $\hat{\gamma}_{x1} = 0.23$ ,  $\hat{\gamma}_{x2} = 0.22$  and  $\hat{\sigma}_u^2 = 0.29$  (true mean number of married people, 24.40; standard deviation (between areas), 13.25);
- (b) larger areas ( $100 < K_i \leq 200$ )— $\hat{\gamma}_0 = 0.78$ ,  $\hat{\gamma}_y = 0.95$ ,  $\hat{\gamma}_{x1} = 0.08$ ,  $\hat{\gamma}_{x2} = 0.25$  and  $\hat{\sigma}_u^2 = 0.49$ , in line with the simulation results (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 non-response perform much better than the other two predictors. Of the two, the proposed predictor  $\hat{M}_{i,NEW}$  has a smaller RMSE, as is the case also in the simulation study (Figs 2, 5 and 8). Note the relatively large values of the coefficients  $\hat{\gamma}_y$  in the two response models, indicating a high degree of informativeness of the non-response. Also note 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. Summarizing remarks

In this paper we propose a general approach for SAE under informative sampling of areas and within areas, and NMAR non-response 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 have been estimated, we consider them as known and follow the approach of Pfeffermann and Sverchkov (2007) of 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 obtained by substituting the unknown response probabilities by their estimators. A simulation study shows good performance of the approach proposed 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 paper considers the case where the models that are 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 methodology proposed in the present paper is under the frequentist approach. As is well known, there is 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 set-up, with appropriate prior distributions for the models' hyperparameters. See Pfeffermann *et al.* (2006) for application of the Bayesian approach for two-level modelling under informative sampling of first- and second-level units.

## Acknowledgements

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

The opinions that are expressed in this paper are of the authors and do not necessarily represent the policies of the US Bureau of Labor Statistics and the Israel Central Bureau of Statistics.

## References

- Cepillini, R., Siniscialco, M. and Smith, C. A. B. (1955) The estimation of gene frequencies in a random mating population. *Ann. Hum. Genet.*, **20**, 97–115.
- Feder, M. and Pfeffermann, D. (2015) Statistical inference under non-ignorable sampling and nonresponse—an empirical likelihood approach. *Preprint*. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton. (Available from <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), p. 236. Toronto: Holt, Rinehart and Winston.
- 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. In *Proc. 6th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1 (eds L. Le Cam, J. Neyman and E. L. Scott), pp. 697–715. Berkeley: University of California Press.
- Pfeffermann, D. (2011) Modelling of complex survey data: why model?; why is it a problem?; how can we approach it? *Surv. Methodol.*, **37**, 115–136.
- Pfeffermann, D. (2013) New important developments in small area estimation. *Statist. Sci.*, **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. *J. Off. Statist.*, **27**, 181–209.

- Pfeffermann, D. and Sverchkov, M. (2007) Small-area estimation under informative probability sampling of areas and within selected areas. *J. Am. Statist. Ass.*, **102**, 1427–1439.
- Pfeffermann, D. and Sverchkov, M. (2009) Inference under informative sampling. In *Sample Surveys: Inference and Analysis* (eds D. Pfeffermann and C. R. Rao), pp. 455–487. Amsterdam: North-Holland.
- Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*, 2nd edn. New York: Wiley.
- Riddles, K. M., Kim, J. K. and Im, J. (2016) A propensity-score adjustment method for nonignorable nonresponse. *J. Surv. Statist. Methodol.*, **4**, 215–245.
- Rivers, D. (2007) Sampling for web surveys. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*
- 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. *Int. Statist. Rev.*, **55**, 279–294.
- Sverchkov, M. (2008) A new approach to estimation of response probabilities when missing data are not missing at random. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 867–874.
- Sverchkov, M. and Pfeffermann, D. (2004) Prediction of finite population totals based on the sample distribution. *Surv. Methodol.*, **30**, 79–92.
- Verret, F., Rao, J. N. K. and Hidioglou, M. A. (2015) Model-based small area estimation under informative sampling. *Surv. Methodol.*, **41**, 333–347.