Statistical Theory and Related Fields

MULTIVARIATE SMALL AREA ESTIMATION UNDER NONIGNORABLE NONRESPONSE

--Manuscript Draft--

Full Title:	MULTIVARIATE SMALL AREA ESTIMATION UNDER NONIGNORABLE NONRESPONSE
Manuscript Number:	TSTF-2019-0001R1
Article Type:	Special Issue Article
Keywords:	Distribution of missing data; Imputation under nonignorable nonresponse; Missing information principle; MSE estimation; NMAR nonresponse.
Abstract:	We consider multivariate small area estimation under nonignorable, not missing at random (NMAR) nonresponse. We assume a response model that accounts for the different patterns of the observed outcomes, (which values are observed and which ones are missing), and estimate the response probabilities by application of the Missing Information Principle (MIP). By this principle, we first derive the likelihood score equations for the case where the missing outcomes are actually observed, and then integrate out the unobserved outcomes from the score equations with respect to the distribution holding for the missing data. The latter distribution is defined by the distribution fitted to the observed data for the respondents and the response model. The integrated score equations are then solved with respect to the unknown parameters indexing the response model. Once the response probabilities have been estimated, we impute the missing outcomes from their appropriate distribution, yielding a complete data set with no missing values, which is used for predicting the target area means. A parametric bootstrap procedure is developed for assessing the mean squared errors (MSE) of the resulting predictors. We illustrate the approach by a small simulation study.
Order of Authors:	Danny Pfeffermann
	Michael Sverchkov
Response to Reviewers:	our response has been submitted.

MULTIVARIATE SMALL AREA ESTIMATION UNDER NONIGNORABLE NONRESPONSE

Danny Pfeffermann[1] and Michael Sverchkov[2]

[1] *National Statistician of Israel*; *Professor*: *Hebrew University of Jerusalem, Israel and University of Southampton, UK*. (**)

[2] *Bureau of Labor Statistics*, Washington DC, USA

We consider multivariate small area estimation under nonignorable, not missing at random (NMAR) nonresponse. We assume a response model that accounts for the different patterns of the observed outcomes, (which values are observed and which ones are missing), and estimate the response probabilities by application of the Missing Information Principle (MIP). By this principle, we first derive the likelihood score equations for the case where the missing outcomes are actually observed, and then integrate out the unobserved outcomes from the score equations with respect to the distribution holding for the missing data. The latter distribution is defined by the distribution fitted to the observed data for the respondents and the response model. The integrated score equations are then solved with respect to the unknown parameters indexing the response model. Once the response probabilities have been estimated, we impute the missing outcomes from their appropriate distribution, yielding a complete data set with no missing values, which is used for predicting the target area means. A parametric bootstrap procedure is developed for assessing the mean squared errors (MSE) of the resulting predictors. We illustrate the approach by a small simulation study.

Key words: Distribution of missing data, Imputation under nonignorable nonresponse, Missing information principle, MSE estimation, NMAR nonresponse.

Michael Sverchkov, Bureau of Labor Statistics, Sverchkov.Michael@bls.gov

(**) *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, models and assumptions

Let $\{\mathbf{y}_{ij}, \mathbf{x}_{ij}; i = 1,...,M,\ j = 1,...,N_i\}$ 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 $\mathbf{y}_{ij} = (y_{ij,1},...,y_{ij,K})'$ is the vector of outcome values for unit j in area i and $\mathbf{x}_{ij} = (x_{ij,1},...,x_{ij,L})'$ is a vector of corresponding L covariates. Note that the use of a single vector $\mathbf{x}_{ij}$ for the covariates accommodates situations where in fact different covariates, possibly of different dimension, apply to different observations. We assume that the covariates are known for every unit in the population, from a recent census or some administrative files. Suppose that the outcome values follow the generic two-level population model:

$$\begin{aligned} &\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i^U \overset{ind}{\sim} f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i^U),\ i = 1,...,M,\ j = 1,...,N_i \\ &\mathbf{u}_i^U \overset{ind}{\sim} f(\mathbf{u}_i^U);\ E(\mathbf{u}_i^U) = \mathbf{0} = (0,...,0)',\ V(\mathbf{u}_i^U) = \Sigma^U, \end{aligned} \quad , \tag{1}$$

where $\mathbf{u}_i^U = (u_{i,1}^U,...,u_{i,K}^U)'$ is a *K*-dimensional latent random effect.

In the present article we assume that a noninformative sample has been drawn from the above population, but the observed data is incomplete because of not missing at random (NMAR) nonresponse. By noninformative sampling we mean that the sampling probabilities are not related to the outcome variable of interest after conditioning on the model covariates, such that the conditional distribution of the outcome variable in the sample, given the covariates, is the same as the corresponding distribution in the population from which the sample is taken.

In practice, the observed data in a sample are almost never complete due to non-response. The extent of the non-response may differ from unit to unit within an area, with some units providing all the requested information, while others only providing part of it, with different units answering different questions. And to make matters worse, the non-response is NMAR, that is, the probability of some target component of a unit being missing may depend, at least in part, on the missing target value, as well as the other target values for that unit, whether

observed or missing. See e.g., Equation (10) for a simple example. As a consequence, approaches that ignore the non-response and just use the complete responses or those that model the non-response only as functions of the observed covariates may yield biased small area predictors. See the simulation study in Section 5.

As a practical example, consider the Household Expenditure Survey (HES) carried out by Israel's Central Bureau of Statistics. The survey collects information on socio-demographic characteristics, as well as information on income and expenditure. The sample consists of households selected with equal probabilities by a two-stage sampling design. Three important questions asked in this survey (and in other similar surveys across the world) relate to the salary in each of the three months preceding the month of the interview. Table 1 presents the distribution of the observed response patterns of the three variable in the 2017 survey, with “1” defining response and “0” nonresponse. The first position to the left defines the response regarding the salary in the month preceding the interview, the middle position defines the response regarding the salary 2 months before the interview, and the third position defines the response regarding the salary 3 months ago.

Table 1. Response patterns on 3 salary variables in Israel's HES. 2017.

Res. Pattern	000	001	010	011	100	101	110	111	Total
Count	885	23	14	308	20	9	40	9,664	10,963
Percentage	8.1	0.2	0.1	2.8	0.2	0.1	0.4	88.2	100

Pfeffermann and Sikov (2011) found that the response to salary questions is informative but they did not consider SAE and restricted to a single target variable. For further discussion and illustrations of NMAR nonresponse and related concepts, see, Rubin (1976), Little (1982), Little and Rubin (2002) , Pfeffermann and Sikov (2011), and references therein.

Returning to the present article, the target is to impute the missing data and use the observed and missing data for estimating the small area means, or other summary measures of interest. It may come as a surprise, but we are not familiar

with published articles considering small area estimation under NMAR nonresponse, except for Sverchkov and Pfeffermann (2018), which treats the case of univariate outcomes. The present paper extends the methodology developed in that article. See Pfeffermann and Sikov (2011) and Riddles et al. (2016) for reviews and many references addressing the problem of NMAR nonresponse when fitting models to survey data, but with no attention to SAE applications.

Define the response indicator $R_{ij,k} = 1(0)$ if $y_{ij,k}$ is observed (unobserved), and let $\mathbf{R}_{ij} = (R_{ij,1},...,R_{ij,K})'$.

Assumption 1.

(1a) The response occurs independently between the units,

(1b) $\Pr[\mathbf{R}_{ij} = \mathbf{r} \mid (\mathbf{y}_{i^*j^*}, \mathbf{x}_{i^*j^*}, \mathbf{u}_{i^*}^U), i^* = 1,....M,\ j^* = 1,..., N_i] = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, \mathbf{x}_{ij}]$.

As noted in Sverchkov and Pfeffermann (2018), Assumption 1b is very reasonable. In particular, it states that the probability to respond to the target variable $\boldsymbol{y}_{ij}$ does not depend on the corresponding random effect given $\boldsymbol{y}_{ij}$, $\Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, u_i^U, \mathbf{x}_{ij}] = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, \mathbf{x}_{ij}]$. Furthermore, it guarantees the identification of the response model. See Remark 3 in Section 2 for further discussion.

Note that under (1) and Assumption 1,

$$\begin{aligned} &f[\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i^U, \mathbf{R}_{ij}, \{(\mathbf{y}_{i^*j^*}, \mathbf{x}_{i^*j^*}, \mathbf{R}_{i^*j^*}, \mathbf{u}_{i^*}^U), i^* = 1...M,\ j^* = 1...N_i; (i^*, j^*) \neq (i, j)\}] \\ &= f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i^U, \mathbf{R}_{ij}). \end{aligned} \tag{2}$$

We assume a parametric form for the completely observed outcomes,

$$\begin{aligned} &\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1} = (1,...,1)' \sim f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i; \theta_1) = f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}; \theta_1); \\ &\mathbf{u}_i = \mathbf{u}_i^U - E(\mathbf{u}_i^U \mid \mathbf{R}_{ij} = \mathbf{1}) \sim f_R(\mathbf{u}_i; \theta_2) = f(\mathbf{u}_i \mid \mathbf{R}_{ij} = \mathbf{1}; \theta_2),\ E_R(\mathbf{u}_i; \theta_2) = 0. \end{aligned} \tag{3}$$

Note that in general, $\mathbf{u}_i^U$ and $\mathbf{u}_i$ are different if the nonresponse is NMAR.

Assumption 2. The subset $\{(i, j): \mathbf{R}_{ij} = \mathbf{1}\}$ is not empty for every sampled area, such that the parameters $\theta = (\theta_1, \theta_2)$ can be estimated by restricting to the fully observed data (units with no missing data), using classical small area estimation (SAE) procedures.

Remark 1. Assumption 2 is for convenience and it is sufficient for our present approach to have fully observed data in only sufficient number of areas to allow efficient estimation of the parameters $\theta = (\theta_1, \theta_2)$. Additionally, for a general response model under which the response to any given component of the multivariate target variable $\boldsymbol{y}$ may depend on the component itself as well as the other components, with possibly different coefficients for each component, (see for example Equation 10 in Section 4), we also require sufficient number of observations for each response pattern $\boldsymbol{R_{ij}}$, thus allowing efficient estimation of the response model for each component.

Denote by $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ the estimate of θ obtained that way. For known θ, the best predictor of the random effect $\boldsymbol{u}_i$ given the completely observed data, $O_C = \{(\mathbf{y}_{ij} : \mathbf{R}_{ij} = \mathbf{1}), \mathbf{x}_{ij}, i = 1,..., M,\ j = 1,..., N_i\}$, is $E(\mathbf{u}_i \mid O_C; \theta)$. We predict, $\hat{\mathbf{u}}_i = E(\mathbf{u}_i \mid O_C; \theta = \hat{\theta})$.

Our proposed procedure to deal with the multivariate informative (NMAR) nonresponse consists of the following steps:

1- Fit a parametric model for the completely observed outcomes, (Equation 3).

2- Fit an appropriate parametric model for the response probabilities, which may depend on the outcome and the covariates (Assumption 1b), indexed by the unknown vector parameter γ; $p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma) = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma]$, with $p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)$ differentiable with respect to γ. See Section 2 for details.

2- Impute the missing outcomes from their appropriate distribution with the unknown parameters $(\theta_1, \theta_2, \gamma)$ replaced by their sample estimates, and then use the "complete" sample data (observed and imputed values), to predict the small

area means or other area measures of interest. See Section 3 for the imputation equations under the model. Since we assume noninformative sampling such that if there was no nonresponse, the sample data would follow the same model as in the population, in what follows we do not distinguish between the population and sample data and consider the population data as our sample. The results of the present study can easily be generalized to the case where first a sample is selected from the finite population by some non-informative or informative sampling scheme, and then nonresponse occurs. In this case one can use the estimated distribution (3) and the estimated response model for imputation of the missing sample data as defined in the present article. Once the missing sample data are imputed, the small area means of interest can be estimated using the approach of Pfeffermann and Sverchkov (2007).

In the next section we apply the MIP principle for estimating the response model parameters and discuss some related questions. In Section 3 we develop the imputation equations for the missing data, which, when combined with the observed data, permit simple estimation of the small area means or other area parameters of interest. In Section 4 we propose a parametric bootstrap procedure for estimating the prediction Root MSE of the resulting predictors. We illustrate our approach with a small simulation study in Section 5 and conclude with a summary of the main outcomes in Section 6.

2. Estimation of response model parameters

If the missing outcome values were actually observed, the vector parameter γ, indexing the response probabilities model, could be estimated by solving the likelihood equations:

$$\sum_{\mathbf{r}=(0,...,0)'}^{(1,...,1)'} \sum_{(i,j):\mathbf{R}_{ij}=\mathbf{r}} \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} = 0, \qquad (4)$$

where the external summation is over all the K-dimension vectors with 0,1 elements.

In practice, the missing data are unobserved for $\mathbf{R}_{ij} \neq \mathbf{1}$ and hence the likelihood equations (4) are not operational. However, one may apply in this case the missing information principle (MIP; Cepillini et al. 1955, Orchard and Woodbury, 1972). See, in particular, Sverchkov (2008), Sverchkov and Pfeffermann (2018), and Riddles et al. (2016) for recent applications of the principle to handle univariate NMAR nonresponse.

Missing Information Principle: Let $O = \{(y_{ij.k} : R_{ij,k} = 1), \mathbf{x}_{ij}, i = 1,...,M,\ j = 1,...,N_i\}$ denote all the observed data. Since no observations are available for elements $(ij,k) : R_{ij,k} = 0$, solve instead the best predictor of (4) given the observed data:

$$E\left(\sum_{\mathbf{r}=(0,...,0)'}^{(1,...,1)'} \sum_{(i,j):\mathbf{R}_{ij}=\mathbf{r}} \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} \middle| O \right) =$$
$$E\left[\sum_{\mathbf{r}=(0,...,0)'}^{(1,...,1)'} \sum_{(i,j):\mathbf{R}_{ij}=\mathbf{r}} E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} \middle| O, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{r} \right) \middle| O \right] = 0. \quad (5)$$

The expectation $E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} \middle| O, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{r} \right)$ can be approximated and solved as follows: Let α denote the set of indexes with observed values $\mathbf{y}_{ij,k}$ and β denote the complement of α, i.e., $\mathbf{y}_{ij,\alpha} = \{y_{ij,k}; r_k = 1\}$, $\mathbf{y}_{ij,\beta} = \{y_{ij,k}; r_k = 0\}$. Denote, $\mathbf{R}_{ij,\alpha} = (R_{ij,k} : k \in \alpha)$, $\mathbf{R}_{ij,\beta} = (R_{ij,k} : k \in \beta)$ and define by $\mathbf{1}_\beta, \mathbf{1}_\alpha$ the corresponding unit vectors of respective dimensions. By Assumption (1b),

$$E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} \middle| O, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{r} \right) =$$
$$\int \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{r})\, d\,\mathbf{y}_{ij,\beta} =$$
$$= \int \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma} \times$$

$$\times \frac{\{[\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta \mid \mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij})]^{-1}-1\} f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1}) d\mathbf{y}_{ij,\beta}}{\int [\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta \mid \mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij})]^{-1} f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1}) d\mathbf{y}_{ij,\beta} - 1}; \quad (6)$$

$$\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta \mid \mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij}) = \frac{p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij};\gamma)}{\int p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij};\gamma) f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1}) d\mathbf{y}_{ij,\beta}}.$$

Finally, solve (5) with respect to γ by substituting $f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1};\hat{\theta}_1)$ $= \frac{f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij},\mathbf{u}_i;\hat{\theta}_1)}{\int f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij},\mathbf{u}_i;\hat{\theta}_1) d\mathbf{y}_{ij,\beta}}$ for $f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1})$, replacing $\mathbf{u}_i$ by $\hat{\mathbf{u}}_i$ and dropping the external expectation. See Sverchkov and Pfeffermann (2018) for a similar approximation in the univariate case.

The last equality (product) in (6) extends to the multivariate case the following fundamental relationship between the sample and sample-complement distributions, derived in Sverchkov and Pfeffermann (2004) for the univariate case:

$$f(y_{ij} \mid x_{ij},u_i,R_{ij}=0) = \frac{[p_r^{-1}(y_{ij},x_{ij})-1] f(y_{ij} \mid x_{ij},u_i,R_{ij}=1)}{E\{[p_r^{-1}(y_{ij},x_{ij})-1] \mid x_{ij},u_i,R_{ij}=1\}}. \quad (7)$$

Equation (7) and its multivariate extension in Equation (6) form the basis for our proposed approach. It states that the distribution of an unobserved (missing) value y_{ij} is defined mathematically by the distribution of y_{ij} if it was observed, and the response model. Notice that under NMAR nonresponse, the distribution of y_{ij} given that the unit responded is different from the distribution of y_{ij} given that the unit did not respond, and also different from the population distribution of y_{ij}, before nonresponse takes place. The proof of the multivariate extension applied in (6) follows the same simple steps of the proof of (7) in Sverchkov and Pfeffermann (2004), utilizing Bayes theorem. See also Sverchkov (2008) and Riddles et al. (2016).

In the Appendix, we illustrate the construction of Equation (6) under the mixed logistic model for the outcome variable.

Remark 2. The dimension of the set of equations in (5) is equal to the dimension of γ indexing the response model and hence it is impossible to estimate the parameters γ and the parameters $\theta = (\theta_1, \theta_2)$ of the outcome model defined by (3), by solely solving this set.

Remark 3. A fundamental question regarding the use of the MIP equations (5) is the existence of a unique solution, or more generally, the identifiability of the response model. For the univariate case, Riddles et al. (2016) deal with NMAR nonresponse in the general context of sample surveys by following an approach proposed by Sverchkov (2008), which is similar to our present approach. Riddles et al. (2016) 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. Covariates 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 effects, given the outcome value, (which depends on the random effects). In practice, the random effects are unobservable but we estimate them and then solve the equations (5) by conditioning on the estimated effects. So, it is actually the estimated random effects that play the role of the covariates $\mathbf{x}_2$. In practice, other covariates that are predictive of the outcome but not of the response might exist as well.

3. Imputation of the missing data.

Once the parameters θ and γ are estimated, the estimates can be substituted (together with $\hat{\mathbf{u}}_i$) into the model holding for the missing data, using the relationship used in (6), yielding the following estimated distribution. Let $\mathbf{y}_{ij,\beta} = \{y_{ij,k}; r_k = 0\}$ define, as before, the unobserved data.

$$f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \hat{\mathbf{u}}_i, \mathbf{R}_{ij} = \mathbf{r}; \hat{\gamma}, \hat{\theta}) =$$

$$\frac{\left[\left(\frac{p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \hat{\gamma})}{\int p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \hat{\gamma}) f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \hat{\mathbf{u}}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta}}\right)^{-1} - 1\right] \frac{f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \hat{\mathbf{u}}_i; \hat{\theta}_1)}{\int f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \hat{\mathbf{u}}_i; \hat{\theta}_1) d\mathbf{y}_{ij,\beta}}}{\int \left(\frac{p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \hat{\gamma})}{\int p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \hat{\gamma}) f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \hat{\mathbf{u}}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta}}\right)^{-1} \frac{f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \hat{\mathbf{u}}_i; \hat{\theta}_1)}{\int f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \hat{\mathbf{u}}_i; \hat{\theta}_1) d\mathbf{y}_{ij,\beta}} d\mathbf{y}_{ij,\beta} - 1}. \quad (8)$$

Note again that the distribution $f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \hat{\mathbf{u}}_i; \hat{\theta}_1)$ is of the observed data and can thus be estimated from the data using standard SAE model fitting procedures.

Imputation of the missing data can be carried out by drawing at random from the distribution (8). One may draw a single observation or multiple observations.

Once the missing observations are imputed, prediction of the true population mean of the outcome variable or other measures of interest is carried out by application of standard procedures. See the empirical study in Section 5.

Remark 4. By Assumption 1, the response occurs independently between units.

4. Estimation of Prediction MSE

As in any other statistical inference problem, one has to assess the error of the resulting predictors. In SAE applications under the frequentist paradigm, it is common to estimate the Root Prediction Mean Squared Error (RPMSE). It is quite obvious that no analytic expression of the RPMSE can be derived, given the complexity of the prediction procedure, and we therefore propose a bootstrap procedure. As before, we assume for convenience no sampling, such that the

sample consists of all the population units. See Remark 5 below. The proposed bootstrap procedure consists of the following steps:

B0- Impute the missing values as developed in Section 3. Consider the pseudo-population of complete responses as the "true" population and calculate the corresponding true-pseudo area means.

B1- For each unit (i, j) with complete observation $\boldsymbol{y}_{ij}^{c}$ generated in Step B0, draw observed outcomes with probabilities $p_{\mathbf{r}}(\mathbf{y}_{ij}^{c}, \mathbf{x}_{ij}; \hat{\gamma})$.

B2- Apply all estimation and imputation procedures described in Sections 2 and 3 to the observed sample obtained in Step B1. Estimate all the area means.

B3- Repeat Steps B1 and B2 independently B times (B large) and compute for each area i the bootstrap RPMSE,

$$RPMSE_{\mathrm{m},k} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_{\mathrm{m},k,b} - \bar{Y}_{\mathrm{m},k}^{\mathrm{B0}})^2;\ m = 1,...,\mathrm{M}, b = 1,...\mathrm{B}, \qquad (9)$$

where $\hat{\bar{Y}}_{\mathrm{m},k,b}$ is the predictor obtained from bootstrap sample b for the mean of the k-th component of the outcome variable in area m and $\bar{Y}_{\mathrm{m},k}^{\mathrm{B0}}$ is the corresponding pseudo mean in area m as obtained in Step B0.

Remark 5. The bootstrap procedure outlined above is partly design-based in the sense that we consider a single pseudo population and the models are used only for estimating the response probabilities and the model holding for the completely observed data. The procedure can easily be extended in two ways. First, we may generate a new pseudo population for each bootstrap sample, thus accounting also for the variability induced by the random generation of the population values. Second, we may extended the procedure to the case where a sample is selected from the population and nonresponse occurs in the sample, by first obtaining complete sample observations as in Step B0 and then generating a pseudo population using the procedure of Sverchkov and Pfeffermann (2004). Thereafter, a sample is drawn from the pseudo population with the same sampling design that was used for drawing the original sample. The other steps

follow Steps B1-B3 above (with or without accounting for the generation of the pseudo population, i.e., by generating only one pseudo population or generating a new population each time).

5. Simulation Study

In this section we describe the results of a simulation experiment when applying the procedures proposed in Sections 2, 3 and 4 (assuming no sampling and a single pseudo population).

The experiment consists of the following steps:

S1- *Generation of population values*: generate for each area $i, i = 1,...,300$ and for each unit $j, j = 1,...,50$ binary covariate values x_{ij} with $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0) = 0.5$, random effects $\mathbf{u}_i = (u_{i,1}, u_{i,2})' \sim N(\mathbf{0}, \mathbf{I})$, $i = 1,...,300$, and corresponding independent outcome values from the mixed logistic model**,**

$$
\begin{aligned}
p_{y_1}(x_{ij}, \mathbf{u}_i) &= \Pr(y_{ij,1} = 1 \mid x_{ij}, \mathbf{u}_i) \\
&= \exp(-.1 - x_{ij} + u_{i,1}) / [1 + \exp(-.1 - x_{ij} + u_{i,1})], \\
p_{y_2}(x_{ij}, \mathbf{u}_i) &= \Pr(y_{ij,2} = 1 \mid x_{ij}, \mathbf{u}_i) = \exp(.9 + u_{i,2}) / [1 + \exp(.9 + u_{i,2})]. \qquad (9)
\end{aligned}
$$

Remark 6. The random effects are generated independently but they are not assumed to be independent in the estimation process.

S2- *Response mechanism*: compute response probabilities for unit j in area i as:

$$
p_{\mathbf{r}}(\mathbf{y}_{ij}, x_{ij}, \gamma) = \begin{cases} C(x_{ij}, \mathbf{y}_{ij}) \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij,1} + \gamma_3 y_{ij,2}), & \text{if} \quad \mathbf{r} = (1,1)' \\ C(x_{ij}, \mathbf{y}_{ij}) \exp(\gamma_4 + \gamma_5 x_{ij} + \gamma_6 y_{ij,1} + \gamma_7 y_{ij,2}), & \text{if} \quad \mathbf{r} = (1,0)' \\ C(x_{ij}, \mathbf{y}_{ij}) \exp(\gamma_8 + \gamma_9 x_{ij} + \gamma_{10} y_{ij,1} + \gamma_{11} y_{ij,2}), & \text{if} \quad \mathbf{r} = (0,1)' \\ C(x_{ij}, \mathbf{y}_{ij}), & \text{if} \quad \mathbf{r} = (0,0)' \end{cases}; \quad (10)
$$

$$
\begin{aligned}
C(x_{ij}, \mathbf{y}_{ij}) = [1 &+ \exp(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij,1} + \gamma_3 y_{ij,2}) + \exp(\gamma_4 + \gamma_5 x_{ij} + \gamma_6 y_{ij,1} + \gamma_7 y_{ij,2}) \\
&+ \exp(\gamma_8 + \gamma_9 x_{ij} + \gamma_{10} y_{ij,1} + \gamma_{11} y_{ij,2})]^{-1};
\end{aligned}
$$

$\gamma_0 = 0,\ \gamma_1 = -.5,\ \gamma_2 = 3,\ \gamma_3 = -3,\ \gamma_4 = 0,\ \gamma_5 = -.5,\ \gamma_6 = 2,\ \gamma_7 = -2,\ \gamma_8 = 0,\ \gamma_9 = -.5,$ $\gamma_{10} = 1,\ \gamma_{11} = -1$. Clearly, the nonresponse is NMAR since the response probabilities depend on the outcomes. Notice that the response for $y_{ij,1}, y_{ij,2}$ is generated independently between units.

Remark 7. We generated a single (finite) population and hence, a single set of response probabilities.

S3- *Generating responses*: generate responses from the (single) population generated in S1, with response probabilities defined in S2 (Equation 10).

S4- *Fitting respondents' model*: estimate $\hat{p}_{y_1}(x_{ij}, \mathbf{u}_i) = \hat{\Pr}(y_{ij,1} = 1 | x_{ij}, \hat{\mathbf{u}}_i, \mathbf{R}_{ij} = \mathbf{1})$, $\hat{p}_{y_2}(x_{ij}, \mathbf{u}_i) = \hat{\Pr}(y_{ij,2} = 1 | x_{ij}, \hat{\mathbf{u}}_i, \mathbf{R}_{ij} = \mathbf{1})$ by fitting the mixed logistic model (9), using PROC NLMIX in SAS with default options. Notice that the model (9) is not the true respondents' model under the response model (10), because of the NMAR nonresponse.

S5. *Estimation of response probabilities*: assume the parametric response model (10), compute the expectations in (6) under the estimated models $\hat{p}_{y_1}(x_{ij}, \hat{\mathbf{u}}_i)$, $\hat{p}_{y_2}(x_{ij}, \hat{\mathbf{u}}_i)$ in Step S4 and estimate γ, using the procedure described in Section 2. See Sverchkov and Pfeffermann (2018) for numerical details.

S6. *Imputation of missing data*: impute the unobserved data from the distribution of the missing data defined in Section 3, which in the present case reduces to:

$$f(\mathbf{y}_{ij,\beta} | \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{r})$$
$$= \frac{\{[\Pr(\mathbf{R}_{ij,\beta} = \mathbf{1}_\beta | \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij,\alpha} = \mathbf{1}_\alpha, \mathbf{y}_{ij})]^{-1} - 1\} f(\mathbf{y}_{ij,\beta} | \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta}}{\int [\Pr(\mathbf{R}_{ij,\beta} = \mathbf{1}_\beta | \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij,\alpha} = \mathbf{1}_\alpha, \mathbf{y}_{ij})]^{-1} f(\mathbf{y}_{ij,\beta} | \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta} - 1}.$$

Remark 8. We imputed a single value for each missing value but one may impute several values, using a multiple imputation approach.

Repeat Steps S3-S6 independently 500 times.

***Predictors considered*:** compute the following predictors for each area on each simulation.

1. $\hat{\bar{Y}}_{i,1}^{ign} = N_i^{-1}\{\sum_{j,R_{ij,1}=1} y_{ij,1} + \sum_{k=1,R_{ik,1}=0}^{N_i} \hat{p}_{y_1}(x_{ik},\hat{\mathbf{u}}_i)\}$,

$\hat{\bar{Y}}_{i,2}^{ign} = N_i^{-1}\{\sum_{j,R_{ij,2}=1} y_{ij,2} + \sum_{k=1,R_{ik,2}=0}^{N_i} \hat{p}_{y_2}(x_{ik},\hat{\mathbf{u}}_i)\}$.

The predictors $\hat{\bar{Y}}_{i,1}^{ign}, \hat{\bar{Y}}_{i,2}^{ign}$ ignore the response process and "assume" that the population distribution holds also for the observed outcomes.

2. $\hat{\bar{Y}}_{i,1}^{new} = N_i^{-1}\sum_{j=1}^{N_i} y_{ij,1}^{imp}$, $\hat{\bar{Y}}_{i,2}^{new} = N_i^{-1}\sum_{j=1}^{N_i} y_{ij,2}^{imp}$, where $y_{ij,k}^{imp} = y_{ij,k}$ if $y_{ij,k}$ is observed, and $y_{ij,k}^{imp}$ is the imputed value from Step S6 if $y_{ij,k}$ is missing ($k = 1,2$).

The estimators $\hat{\bar{Y}}_{i,1}^{new}, \hat{\bar{Y}}_{i,2}^{new}$ are our proposed estimators, accounting for the multivariate NMAR nonresponse.

Statistics considered for assessment of the of predictors' performance

Denote by $\bar{Y}_{i,k,r}$ the true mean of area i on the *r*-th simulation (for first or second coordinate, *k*= 1 or 2), and let $\hat{\bar{Y}}_{i,k,r}$ represent the first or second predictors defined above, $r = 1,...,500$.

$$Bias_{i,k} = \frac{\sum_{r=1}^{500}(\hat{\bar{Y}}_{i,k,r} - \bar{Y}_{i,k,r})}{500} \;;\; RPMSE_{i,k} = \frac{\sum_{r=1}^{500}(\hat{\bar{Y}}_{i,k,r} - \bar{Y}_{i,k,r})^2}{500};$$

$$\mathrm{Rel}Bias_{i,k} = \frac{Bias_{i,k}}{\sqrt{V_{i,k}}};\; V_{i,k} = \frac{\sum_{r=1}^{500}(\hat{\bar{Y}}_{i,k,r} - \frac{1}{500}\sum_{r=1}^{500}\hat{\bar{Y}}_{i,k,r})^2}{500};$$

$$\mathrm{RelRPMSE}_{i,k} = \frac{\sqrt{RPMSE_{i,k}}}{\left(\frac{1}{500}\sum_{r=1}^{500}\bar{Y}_{i,k,r}\right)}.$$

We calculated for each area the average (over the 500 simulations) of the number of complete responses and ordered the areas by these averages (the smallest mean number of complete responses is 2.3, the largest is 28.1).

S7. *Estimation of the Root Prediction MSE (RPMSE)*: compute bootstrap estimates of RPMSE following the steps B0-B3 in Section 4.

In the following four figures we show the results for $\mathrm{Rel}Bias_{i,k}$ and $\mathrm{RelRMSE}_{i,k}$, $k=1,2$ for each area, with the areas ordered as above, starting with the area with the smallest number of complete responses.

Figure 1. $\mathrm{Rel}Bias_{i,1}$ of $\hat{\bar{Y}}_{i,1}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,1}^{new}$ ("+")

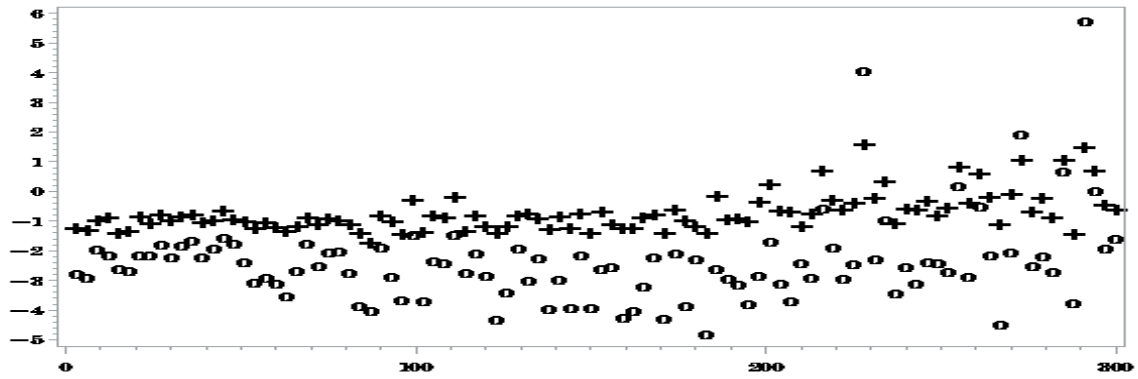
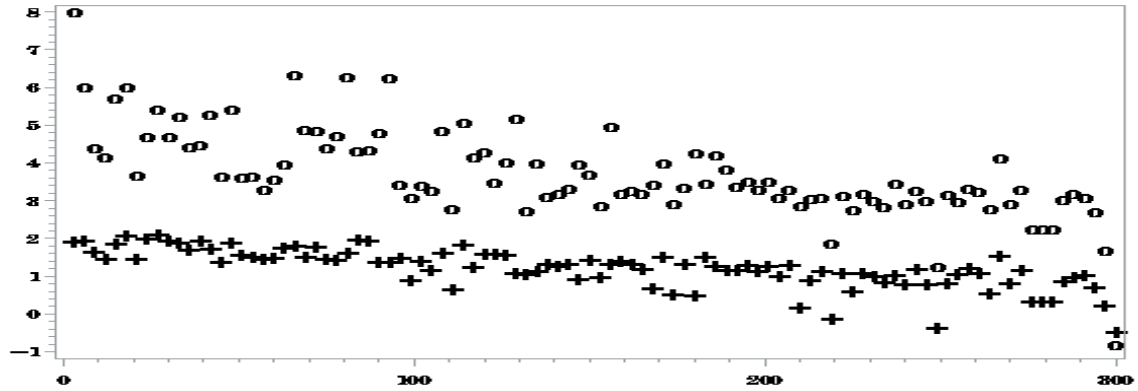


Figure 2. $\mathrm{Rel}Bias_{i,2}$ of $\hat{\bar{Y}}_{i,2}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,2}^{new}$ ("+")



Figures 1 and 2 show how the proposed method reduces very significantly the bias due to NMAR nonresponse. As expected, the bias of both set of predictors decreases as the number of complete responses increases but our proposed predictors are seen to be much less biased.

Figure 3. $\mathrm{RelRPMSE}_{i,1}$ of $\hat{\bar{Y}}_{i,1}^{ign}$ (“o”) and $\hat{\bar{Y}}_{i,1}^{new}$ (“+”)

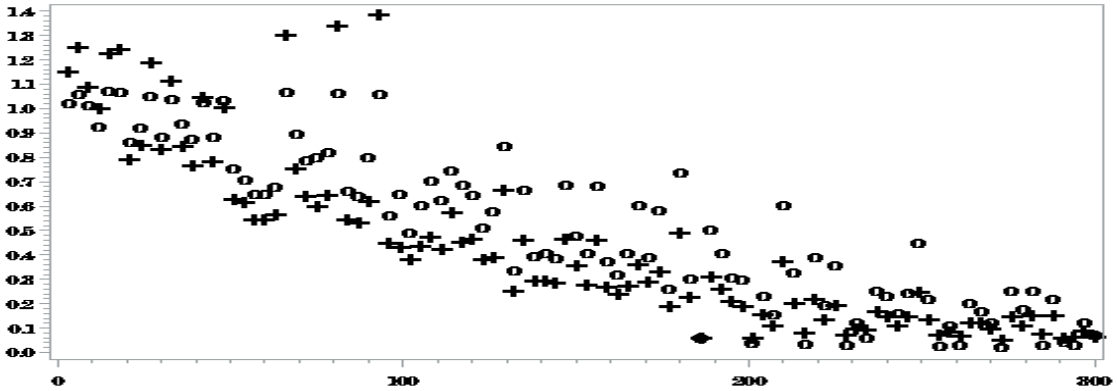
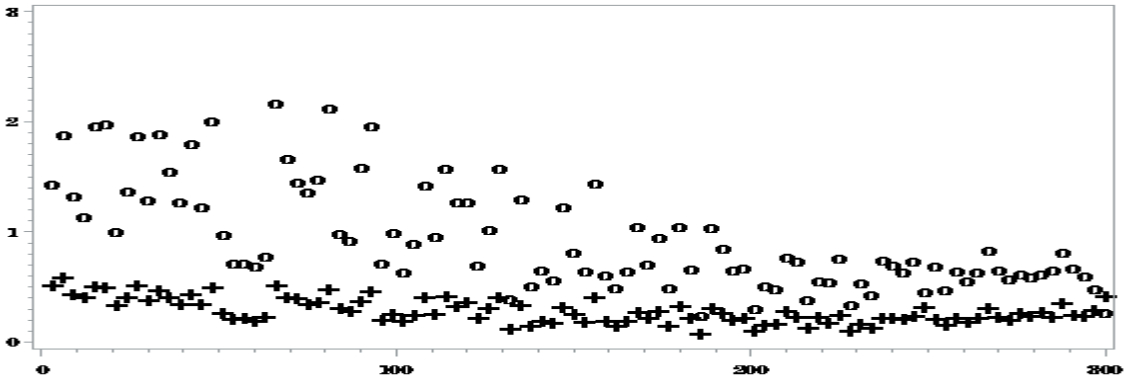


Figure 4. $\mathrm{RelRPMSE}_{i,2}$ of $\hat{\bar{Y}}_{i,2}^{ign}$ (“o”) and $\hat{\bar{Y}}_{i,2}^{new}$ (“+”)



The reduction in *RelRMSE* by accounting for the NMAR nonresponse in Figure 3 is not big, which is explained by the fact that the bias of the predictors that ignore the nonresponse is not very high in this case. Notice in this respect that the average number of missing values $y_{ij,1}$ over the 500 simulations is 5531.5, compared to an average number of 6014.6 missing values of $y_{ij,2}$. Nonetheless, when averaging the $\mathrm{RelRMSE}_{i,1}$ over all the areas we find that,

$$\mathrm{Average}[\mathrm{RelRPMSE}_{i,1}(\hat{\bar{Y}}_{i,1}^{ign})] = \frac{1}{300}\sum_{i=1}^{300}\mathrm{RelRPMSE}_{i,1}(\hat{\bar{Y}}_{i,1}^{ign}) = 0.51,$$

$$\mathrm{Average}[\mathrm{RelRPMSE}_{i,1}(\hat{\bar{Y}}_{i,1}^{new})] = 0.44.$$

When estimating $\bar{Y}_{i,2}$ in Figure 4, the reduction in the $\mathrm{RelRPMSE}$ by use of the proposed procedure is much more drastic, particularly in the areas with small

numbers of complete responses, due to the large bias when ignoring the NMAR nonresponse.

$$\text{Average}[\text{RelRPMSE}_{i,2}(\hat{\bar{Y}}_{i,2}^{ign})] = 0.93, \quad \text{Average}[\text{RelRPMSE}_{i,2}(\hat{\bar{Y}}_{i,2}^{new})] = 0.28.$$

Next, we study the sensitivity of the proposed approach to correct specification of the response model. For this, we repeated the same simulation study, but by computing the response probabilities as:

$$p_{\mathbf{r}}(\mathbf{y}_{ij}, x_{ij}, \gamma) = \begin{cases} C(x_{ij}, \mathbf{y}_{ij}) \exp[\gamma_0 + \gamma_1 x_{ij}(\gamma_2 y_{ij,1} + \gamma_3 y_{ij,2})], & \text{if} \quad \mathbf{r} = (1,1)' \\ C(x_{ij}, \mathbf{y}_{ij}) \exp[\gamma_4 + \gamma_5 x_{ij}(\gamma_6 y_{ij,1} + \gamma_7 y_{ij,2})], & \text{if} \quad \mathbf{r} = (1,0)' \\ C(x_{ij}, \mathbf{y}_{ij}) \exp[(\gamma_8 + \gamma_9 x_{ij}(\gamma_{10} y_{ij,1} + \gamma_{11} y_{ij,2})], & \text{if} \quad \mathbf{r} = (0,1)' \\ C(x_{ij}, \mathbf{y}_{ij}), & \text{if} \quad \mathbf{r} = (0,0)' \end{cases}; \qquad (11)$$

$C(x_{ij}, \mathbf{y}_{ij}) = \{1 + \exp[\gamma_0 + \gamma_1 x_{ij}(\gamma_2 y_{ij,1} + \gamma_3 y_{ij,2})] + \exp[\gamma_4 + \gamma_5 x_{ij}(\gamma_6 y_{ij,1} + \gamma_7 y_{ij,2})]$
$+ \exp[\gamma_8 + \gamma_9 x_{ij}(\gamma_{10} y_{ij,1} + \gamma_{11} y_{ij,2})]\}^{-1}$, with the same coefficients as in (10).

With these response probabilities, the number of complete responses in an area (averaged over the 500 simulations), is in the range [9.3, 18.3].

When estimating the response model parameters in Step S5 of the simulation, we still use the model (10) as the working model, so that the model for the response is misspecified, and so is the model estimated for the missing data. (As mentioned before, the model estimated for the completely observed outcomes is also not correct).

Table 2 compares the true response probabilities (Equation 11) with the average of the estimated response probabilities over the 500 simulations under the misspecified response model (Equation 10). Notice that except in a few cases, that averages of the estimated response probabilities under the misspecified model are close to the true response probabilities, already illustrating lack of sensitivity of our proposed approach to correct specification of the response model, although the differences between the true response probabilities and their estimates are occasionally larger for any given sample.

Table 2. True response probabilities, $p_{r,(i,j)}$, and average of estimated response probabilities, $A\hat{p}_{r,(i,j)}$ under misspecified response model, for different response patterns $r_{ij};(i=0,1,\ j=0,1)$.

x	y_1	y_2	$p_{r,(1,1)}$	$p_{r,(1,0)}$	$p_{r,(0,1)}$	$A\hat{p}_{r,(1,1)}$	$A\hat{p}_{r,(1,0)}$	$A\hat{p}_{r,(0,1)}$
0	0	0	.25	.25	.25	.25	.25	.25
0	0	1	.25	.25	.25	.31	.26	.21
0	1	0	.25	.25	.25	.19	.22	.30
0	1	1	.25	.25	.25	.24	.26	.24
1	0	0	.25	.25	.25	.30	.24	.24
1	0	1	.46	.27	.17	.36	.26	.19
1	1	0	.10	.17	.27	.23	.22	.29
1	1	1	.25	.25	.25	.28	.26	.23

Figure 5. $\mathrm{Rel}Bias_{i,1}$ of $\hat{\bar{Y}}_{i,1}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,1}^{new}$ ("+"), response model misspecified

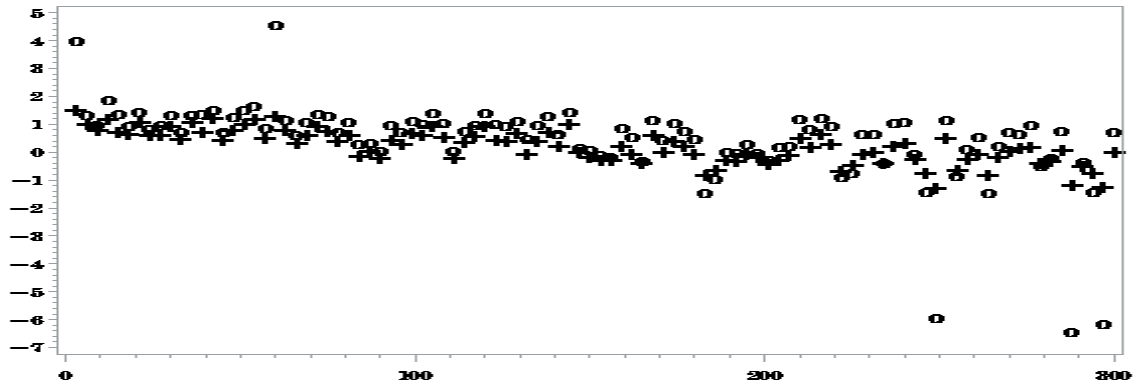
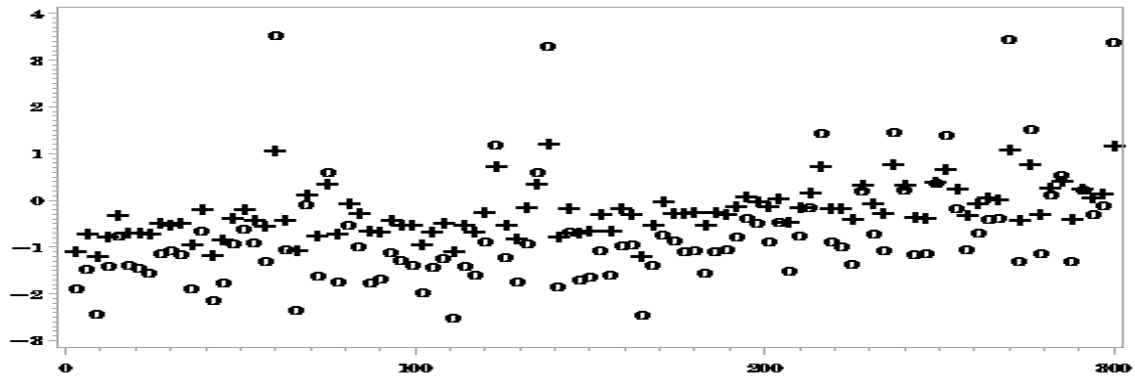


Figure 6. $\mathrm{Rel}Bias_{i,2}$ of $\hat{\bar{Y}}_{i,2}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,2}^{new}$ ("+"), response model misspecified



With the misspecified response model (and the model for the completely observed data), there are no big biases even when ignoring the NMAR nonresponse. Nonetheless, even in this case, when averaging over all the areas,

$$Average[|\,\mathrm{Relbias}\,(\hat{\bar{Y}}_{i,1}^{ign})|]=1.09,\ Average[|\,\mathrm{Relbias}\,(\hat{\bar{Y}}_{i,1}^{new})|]=\ 0.50,$$

$$Average[|\,\mathrm{Relbias}\,(\hat{\bar{Y}}_{i,2}^{ign})|]=\ 1.17,\ Average[|\,\mathrm{Relbias}\,(\hat{\bar{Y}}_{i,2}^{new})|]=\ 0.47.$$

Next we compare the RelRPMSEs of the two estimators.

Figure 7. $\mathrm{RelRPMSE}_{i,1}$ of $\hat{\bar{Y}}_{i,1}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,1}^{new}$ ("+"), response model misspecified

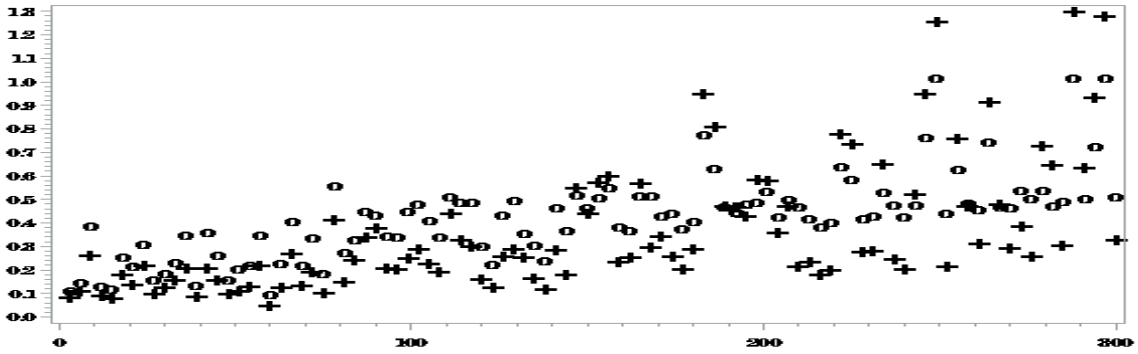
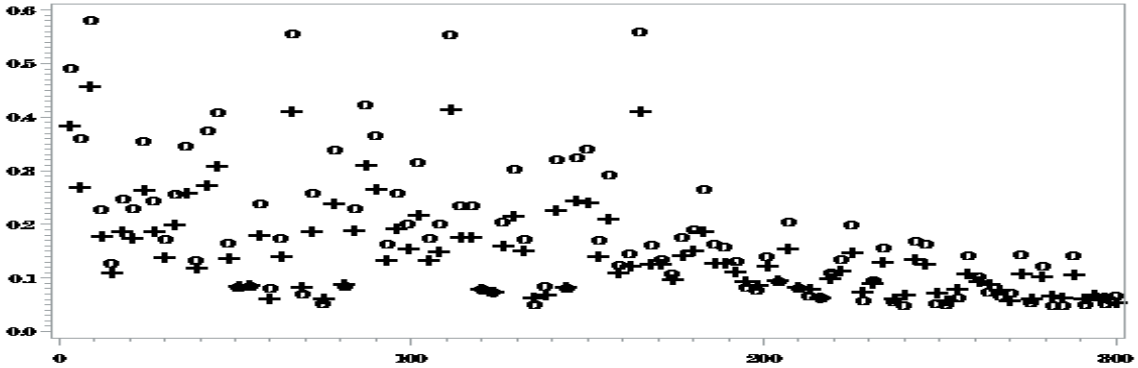


Figure 8. $\mathrm{RelRPMSE}_{i,2}$ of $\hat{\bar{Y}}_{i,2}^{ign}$ ("o") and $\hat{\bar{Y}}_{i,2}^{new}$ ("+"), response model misspecified



Figures 7 and 8 show reduction in the RelRPMSEs when accounting for the NMAR nonresponse in the areas with small number of complete responses. When averaging over all the areas,

$$\mathrm{Average}[\mathrm{RelRPMSE}_{i,1}(\hat{\bar{Y}}_{i,1}^{ign})]=0.41,\mathrm{Average}[\mathit{Rel}\mathrm{RPMSE}_{i,1}(\hat{\bar{Y}}_{i,1}^{new})]=0.34;$$

$$\mathrm{Average}[\mathrm{RelRPMSE}_{i,2}(\hat{\bar{Y}}_{i,2}^{ign})]=0.17,\ \mathrm{Average}[\mathrm{RelRPMSE}_{i,2}(\hat{\bar{Y}}_{i,2}^{new})]=\ 0.14.$$

We conclude that even under the misspecified models, our approach generally yields predictors with smaller *RelRMSEs* than when ignoring the NMAR nonresponse. Clearly, the predictors obtained under this approach have larger variances than when ignoring the NMAR nonresponse, due to all the complex computations involved, so that the large differences in the bias do not always translate into corresponding large differences in the *RelRMSEs*.

Finally, we report the results of RelRPMSE estimation. Due to time limitation, the results so far are based on only 100 parent samples and 50 bootstrap samples for each parent sample. Figures 9 and 10 compare the "true" (empirical) RelRPMSEs over the 100 parent samples, with the mean of the corresponding bootstraps estimates.

Figure 9. $\mathrm{RelRPMSE}_{i,1}$ of $\hat{\bar{Y}}_{i,1}^{new}$ ("+"), and bootstrap estimates ("o")

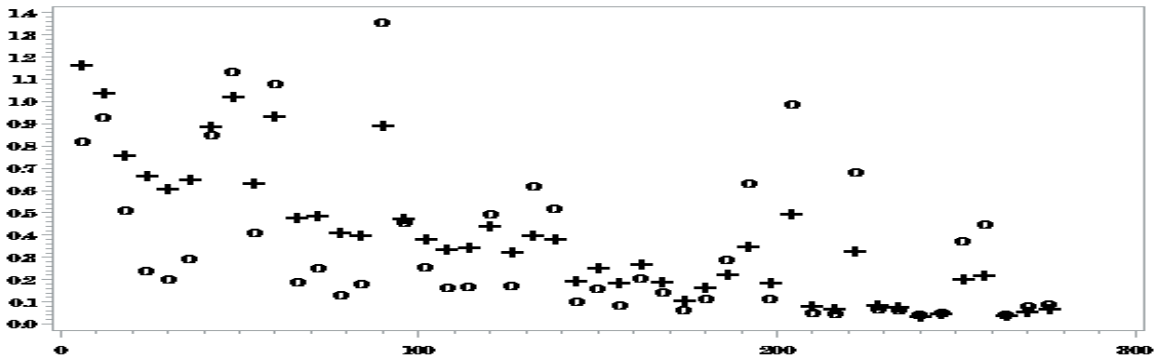
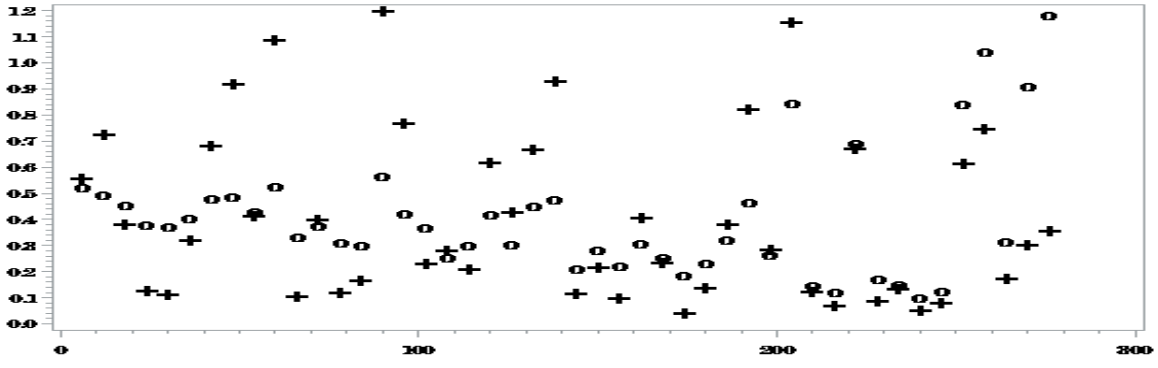


Figure 10. $\mathrm{RelRPMSE}_{i,2}$ of $\hat{\bar{Y}}_{i,2}^{new}$ ("+"), and bootstrap estimates ("o")



The results in Figures 9 and 10 show for most areas good performance of the bootstrap estimators and we believe that with more parent samples and

bootstrap samples, the results will look even better. Even with the current runs, when averaging over all the areas,

$$\text{Average}[\text{RelRPMSE}_{i,1}] = 0.38,\ \text{Average Bootstrap}[\text{RelRPMSE}_{i,1}] = 0.35,$$

$$\text{Average}[\text{RelRPMSE}_{i,2}] = 0.41,\ \text{Average Bootstrap}[\text{RelRPMSE}_{i,2}] = 0.41,$$

illustrating the unbiasedness of the bootstrap estimators when averaging over all the areas.

We compared the empirical RelRPMSE's with the bootstrap estimates also for the case of the misspecified response model and obtained similar results. To save in space, we don't show the corresponding figures.

6. Summary

In this paper we propose a general approach for multivariate SAE under NMAR nonresponse within the selected areas. The approach consists of fitting a model for the observed data and using this model for estimating a postulated multivariate response model by application of the missing information principle. Once the response model is estimated, we derive the model holding for the missing data, which is used for imputing the missing data, thus obtaining a complete file of sample data that is used for estimating the unknown small area parameters. A bootstrap procedure is proposed for estimating the root prediction mean squared errors of the small area predictors, which consists of generating a pseudo population with similar behaviour to the behaviour of the true underlying population, and selecting many samples from the pseudo population and many sets of responses for each sample.

A simulation study shows good performance of our approach in terms of point and RPMSE estimation. The simulation study also illustrates certain robustness to misspecification of the response model. 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 assume general models for the observed data and the response mechanism. Thus, we encourage researchers of SAE to apply the procedure to simulated and real data

sets, with possibly different models assumed for the observed data and the response probabilities.

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.

Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd edn, New York: Wiley.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association* **77**, 237-250.

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. 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.

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

Rubin, D.B. (1976). “Inference and missing data”, *Biometrika*, **63**, 581-590. I

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.

Sverchkov, M., and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random nonresponse. *Journal of the Royal Statistical Society* JRSS-SA, **181**, 981-1008.

Appendix. Illustration of the use of Equation (6) for Estimation of the response probabilities:

Mixed logistic model for the outcome variables with a single covariate.

Consider bivariate variables $\boldsymbol{y_{ij}} = (y_{ij1}, y_{ij2})$, and suppose that the model fitted to the observed data of the respondents is the mixed generalized logistic model,

$$\begin{aligned} p_{y_1}(x_{ij}, \mathbf{u}_i) &= \Pr(y_{ij,1} = 1 \mid x_{ij}, \mathbf{u}_i, \mathbf{R_{ij}} = \mathbf{1}) \\ &= \exp(\alpha_1 + \beta_1 x_{ij} + u_{i,1})[1 + \exp(\alpha_1 + \beta_1 x_{ij} + u_{i,1})]^{-1} \\ p_{y_2}(x_{ij}, \mathbf{u}_i) &= \Pr(y_{ij,2} = 1 \mid x_{ij}, \mathbf{u}_i, \mathbf{R_{ij}} = \mathbf{1}) \\ &= \exp(\alpha_2 + \beta_2 x_{ij} + u_{i,2})[1 + \exp(\alpha_2 + \beta_2 x_{ij} + u_{i,2})]^{-1}, \end{aligned} \tag{8}$$

$$\mathbf{u}_i = (u_{i,1}, u_{i,2})' \sim N(\mathbf{0}, \Sigma).$$

Suppose a generic response model, $p_{\mathbf{r}}(\mathbf{y}_{ij}, x_{ij}; \gamma) = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, x_{ij}; \gamma]$.

We assume that $y_{ij,1}$ and $y_{ij,2}$ are independent given $x_{ij}, \mathbf{u}_i, \mathbf{R_{ij}} = \mathbf{1}$, and that $\Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, x_{ij}, \mathbf{u}_i; \gamma] = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij}, x_{ij}; \gamma]$.

Then, for example, for $\mathbf{r} = (0,1)'$, the components of (6) can be written as,

$$\int \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}; \gamma)}{\partial \gamma}$$

$$\times\{[\Pr(\mathbf{R}_{ij,\beta} = \mathbf{1}_\beta \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij,\alpha} = \mathbf{1}_\alpha, \mathbf{y}_{ij})]^{-1} - 1\} f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta}$$

$$= \frac{\partial \log p_{\mathbf{r}}[(1, y_{ij,2})', \mathbf{x}_{ij}; \gamma]}{\partial \gamma}$$

$$\times\{[\frac{p_{\mathbf{r}}[(1, y_{ij,2})', \mathbf{x}_{ij}; \gamma]}{p_{\mathbf{r}}[(1, y_{ij,2})', \mathbf{x}_{ij}; \gamma] p_{y_1}(x_{ij}, \mathbf{u}_i) + p_{\mathbf{r}}[(0, y_{ij,2})', \mathbf{x}_{ij}; \gamma][1 - p_{y_1}(x_{ij}, \mathbf{u}_i)]}]^{-1} - 1\}$$

$$\times p_{y_1}(x_{ij}, \mathbf{u}_i) + \frac{\partial \log p_{\mathbf{r}}[(0, y_{ij,2})', \mathbf{x}_{ij}; \gamma]}{\partial \gamma}$$

$$\times\{[\frac{p_{\mathbf{r}}[(0,y_{ij,2})',\mathbf{x}_{ij};\gamma]}{p_{\mathbf{r}}[(1,y_{ij,2})',\mathbf{x}_{ij};\gamma]p_{y_1}(x_{ij},\mathbf{u}_i)+p_{\mathbf{r}}[(0,y_{ij,2})',\mathbf{x}_{ij};\gamma][1-p_{y_1}(x_{ij},\mathbf{u}_i)]}]^{-1}-1\}$$
$$\times[1-p_{y_1}(x_{ij},\mathbf{u}_i)],$$

$$\int\{[\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta\,|\,\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij})]^{-1}-1\}\,f(\mathbf{y}_{ij,\beta}\,|\,\mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1})d\mathbf{y}_{ij,\beta}$$

$$=\{[\frac{p_{\mathbf{r}}[(1,y_{ij,2})',\mathbf{x}_{ij};\gamma]}{p_{\mathbf{r}}[(1,y_{ij,2})',\mathbf{x}_{ij};\gamma]p_{y_1}(x_{ij},\mathbf{u}_i)+p_{\mathbf{r}}[(0,y_{ij,2})',\mathbf{x}_{ij};\gamma][1-p_{y_1}(x_{ij},\mathbf{u}_i)]}]^{-1}-1\}$$
$$\times p_{y_1}(x_{ij},\mathbf{u}_i)$$

$$+\{[\frac{p_{\mathbf{r}}[(0,y_{ij,2})',\mathbf{x}_{ij};\gamma]}{p_{\mathbf{r}}[(1,y_{ij,2})',\mathbf{x}_{ij};\gamma]p_{y_1}(x_{ij},\mathbf{u}_i)+p_{\mathbf{r}}[(0,y_{ij,2})',\mathbf{x}_{ij};\gamma][1-p_{y_1}(x_{ij},\mathbf{u}_i)]}]^{-1}-1\}$$
$$\times[1-p_{y_1}(x_{ij},\mathbf{u}_i)].$$

Similar expressions are obtained for $\mathbf{r}=(1,0)'$ and $\mathbf{r}=(0,0)'$.