Imputation under informative sampling

Imputation under informative sampling

Emily Berg * J.K. Kim † Chris Skinner ‡

February 16, 2016

Abstract

Imputed values in surveys are often generated under the assumption that the sampling mechanism is non-informative (or ignorable) and the study variable is missing at random (MAR). When the sampling design is informative, the assumption of MAR in the population does not necessarily imply MAR in the sample. In this case, the classical method of imputation using a model fitted to the sample data does not in general lead to unbiased estimation. To overcome this problem, we consider alternative approaches to imputation assuming MAR in the population. We compare the alternative imputation procedures through simulation and an application to estimation of mean erosion using data from the Conservation Effects Assessment Project.

Key Words: Fractional imputation, Multiple imputation, Missing at random, Variance estimation.

*Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A

†Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A

‡Department of Statistics, The London School of Economics and Political Science, London, WC2A 2AE, U.K.

1 Introduction

Imputation is widely used to handle item nonresponse in surveys. Imputed values are often obtained by fitting parametric regression models relating the variable with missing values to covariates observed for all sample units. The sampling scheme is typically ignored when fitting these models and constructing the imputed values (e.g. Rubin, 1987, sect. 3.6). In this paper, we investigate approaches to imputation where sampling is non-ignorable. We suppose that the non-ignorability arises because sampling is informative, that is sample inclusion is not independent of the variable which is missing given the observed covariates (Pfeffermann, 1993, 2011; Fuller, 2009, ch. 6).

A conventional assumption used to ensure the approximate unbiasedness of the imputed estimator is that values are missing at random (MAR) given the values of the covariates (e.g. Seaman et al., 2013). When sampling is informative, models applying to the population may not apply to the sample and we argue that it is important to distinguish the notions of missing at random in the sample (SMAR) and missing at random in the population (PMAR). If one is willing to assume SMAR then, by appropriately conditioning imputation on sample inclusion, the sampling scheme can be ignored in the construction of the imputed values. In this paper, we suppose that it is only reasonable to assume PMAR.

PMAR may be a more natural assumption than SMAR if the mechanisms underlying the response propensity are conceptualized as inherent characteristics of the units in the population. This perspective might garner support if expert knowledge is available about the missingness process from other surveys, which may employ different sampling schemes. In this case, the knowledge about the missingness process needs to be free of the sample design if this evidence is to be transportable to the survey of interest, so viewing the missingness mechanism as a function of the population characteristics alone is the more natural approach. Similarly, if the missingness mechanism is viewed as a process amenable to scientific examination (e.g. Schafer, 1997, sect. 2.4) then it might be argued, as in the literature on survey analysis

(e.g. Skinner, Holt and Smith, 1989), that it is natural to define and examine such mechanisms in terms of population models rather than sample models. Applying the SMAR assumption to the specific survey of interest would lack credibility if the analyst adopts these perspectives on the response mechanism.

The multiple imputation literature recognizes that complex sampling schemes can affect inference and, in particular, induce bias (Kott, 1995; Reiter et al., 2006). The usual recommendation in this case is to augment the imputation model by including design information, such as clustering and stratification indicators and sample design weights in the covariates (Rubin, 1996; Schenker et al, 2006). Augmenting the imputation model using design information might be expected to make it more likely for SMAR to hold. Conditioning on design weights has been shown to overcome some effects of informative sampling (Rubin, 1996) and we shall consider it as one approach in this paper. However, we shall find that it does not ensure that SMAR holds when PMAR holds and that it does not ensure that the usual imputed estimator is approximately unbiased in the general case of PMAR. Seaman et al. (2012) and Carpenter and Kenward (2013, Ch. 11) have considered other approaches to combining multiple imputation and survey weighting. Their focus is somewhat different, however. They assume conditions under which the usual imputed estimator is approximately unbiased and focus more on issues of multivariate missingness, bias in the multiple imputation variance estimator and the effect of misspecification of the imputation model. Given the potential bias of an approach which conditions on design weights, we shall also consider an alternative design-weighted approach which is widely used for fitting regression models under informative sampling (Pfeffermann, 2011).

The survey sampling literature considers imputation in different inferential frameworks. Inference in the nonresponse model framework (Haziza, 2009) does not depend upon the imputation model and thus avoids the kinds of biases arising from informative sampling considered so far. However, inference does depend upon stronger assumptions about the nonresponse mechanism than a MAR-type assumption and this approach will not be considered further here. The literature adopting an imputation model approach, such as Särndal (1992), Deville and Särndal (1994) and Kim

and Rao (2009), is closer to the approach adopted in this paper but has generally seemed to make assumptions, e.g. Condition 4 of Chauvet et al. (2011), which remove the bias effect of informative sampling. The ideas in this paper are potentially applicable to the methods in this imputation model literature.

In section 2, we consider approaches to imputation and associated assumptions, including, in particular, the distinction between PMAR and SMAR. In the following sections we extend the theory to fractional and multiple imputation frameworks. A limited simulation study then provides evidence on the relative performance of different approaches. An illustration with data from the Conservation Effects Assessment Project, a survey designed to collect information related to water and wind erosion from crop fields, exemplifies a situation in which the data support the use of the survey weights in estimating the imputation model.

2 Framework, Assumptions and Single Imputation

To formalize the problem, assume that the finite population $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i)\,; i \in U_N\}$ with $U_N = \{1, \cdots, N\}$ is a random sample from an infinite population ζ with joint density $f\,(y \mid \boldsymbol{x})\,g\,(\boldsymbol{x})$, the conditional density $f\,(y \mid \boldsymbol{x})$ and the marginal density $g\,(\boldsymbol{x})$. The marginal density $g(\boldsymbol{x})$ is completely unspecified. From a realized finite population, we select a sample $A \subset U_N$ by a probability sampling design. Let I_i be the indicator function of sample selection for unit i, that is, $I_i = 1$ if unit i is selected for the sample and $I_i = 0$ otherwise. From the sample, we collect information about $(\boldsymbol{x}_i, y_i)$, where y_i is the variable of interest and $\boldsymbol{x}_i$ is a vector of auxiliary variables. Let R_i be the indicator function of response on y_i so that we observe y_i if $R_i = 1$ and not if $R_i = 0$. We observe $\boldsymbol{x}_i$ for all sample units. We assume that R_i is defined throughout the finite population, following the stable response assumption of Rubin (1987) or the extended definition of nonresponse used in Fay (1992) and Shao and Steel (1999). We extend the earlier infinite population assumption to suppose that the $(y_i, R_i, I_i, \boldsymbol{x}_i); i \in U_N$ are identically distributed as $(y, R, I, \boldsymbol{x})$.

We are interested in estimating $\theta = \sum_{i=1}^{N} y_i$, the population total of y, or some other function of the finite population values. Assume that the first order inclusion

probability $\pi_i = Pr(I_i = 1)$ is available throughout the sample and so we could use $\hat{\theta}_n = \sum_{i=1}^{N} I_i \pi_i^{-1} y_i$ to estimate θ if y_i were observed throughout the sample. In our case, where y_i is only observed if $R_i = 1$, we can estimate θ using a single imputation approach by setting

$$\hat{\theta}_I = \sum_{i=1}^{N} I_i R_i \pi_i^{-1} y_i + \sum_{i=1}^{N} I_i (1 - R_i) \pi_i^{-1} y_i^*, \tag{1}$$

where y_i^* is the imputed value for y_i. A conventional rationale to achieve approximately unbiased imputed estimation is to generate y_i^* which satisfy

$$E\left\{y_i - y_i^* \mid \boldsymbol{x}_i, I_i = 1, R_i = 0\right\} = 0. \tag{2}$$

To achieve condition (2), we should like to generate imputed values from the conditional distribution $f(y_i \mid \boldsymbol{x}_i, I_i = 1, R_i = 0)$ and, for this purpose, we often assume that

$$f\left(y \mid \boldsymbol{x}, I = 1, R = 1\right) = f\left(y \mid \boldsymbol{x}, I = 1, R = 0\right) \tag{3}$$

and generate imputed values from $f(y_i \mid \boldsymbol{x}_i, I_i = 1, R_i = 1)$, which can be estimated from the observed data. Condition (3) is the usual missing at random (MAR) assumption, as in the formulation of Little (2003), but to emphasize that it depends on the realized sample (i.e. is conditional on $I = 1$) we refer to it as sample missing at random (SMAR). Using the notation $\perp$ from Dawid (1979) to denote (conditional) independence, this condition may alternatively be expressed as

$$y \perp R \mid \boldsymbol{x}, I = 1 \tag{4}$$

and contrasted with

$$y \perp R \mid \boldsymbol{x} \tag{5}$$

which we refer to as population missing at random (PMAR), as discussed earlier. In this paper, we consider approaches to imputation when PMAR holds but SMAR does not. The following lemma identifies properties of the sampling or response mechanisms for which these circumstances do not apply.

Lemma 1 *If PMAR holds, sufficient conditions for SMAR to hold also are either*

1. $I \perp Y \mid \boldsymbol{x}, R$ *or*

2. $R \perp (y, I) \mid \boldsymbol{x}$

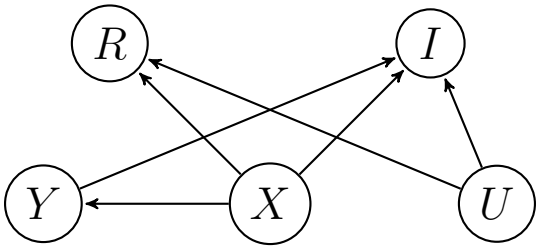
Proof. When condition 1 holds, $f(y \mid \boldsymbol{x}, I = 1, R)$ reduces to $f(y \mid \boldsymbol{x}, R)$ which reduces to $f(y \mid \boldsymbol{x})$ under PMAR. Hence $f(y \mid \boldsymbol{x}, I = 1, R)$ is free of R and SMAR holds. SMAR follows from condition 2 by Lemma 4.2 of Dawid (1979). ■

The first condition states that the sampling mechanism is non-informative given $\mathbf{x}$ (Pfeffermann, 1993; Pfeffermann and Sverchkov, 1999) within both the responding and nonresponding subpopulations. The second condition states that the response mechanism is unrelated to either y or sample inclusion given $\mathbf{x}$.

In general, however, PMAR will not imply SMAR, as is illustrated using the simple example of a population of size 1000 in Table 1, where y is binary and $\mathbf{x}$ is suppressed for simplicity. Taking the empirical proportions to represent probabilities, we see that PMAR holds in the sense that $P(R = 1 \mid y = 0) = P(R = 1 \mid y = 1) = 0.5$ but that SMAR does not hold since $P(R = 1 \mid y = 0, I = 1) = 0.8$ differs from $P(R = 1 \mid y = 1, I = 1) = 0.2$. This effect arises from a three-way association between R, I and y, since we observe that all two-way associations are zero. Thus, not only does PMAR hold, so that $y \perp R$, but also sampling is non-informative, in the sense that $I \perp y$, since $P(I = 1 \mid y = 0) = P(I = 1 \mid y = 1) = 0.2$ and response is unconfounded with sampling (Rubin, 1987) in the sense that $R \perp I$, since $P(R = 1 \mid I = 0) = P(R = 1 \mid I = 1) = 0.5$.

In our simulation study, we shall give a further illustration of how PMAR may hold but SMAR does not. For the simulation, in addition to $(y, R, I, \boldsymbol{x})$, the population contains a latent variable u that is never observed or is unidentified. An example of u may be a design variable that is unavailable to the analyst at the estimation stage. The latent u may introduce correlation in the conditional joint distribution of (y, R, I) given the auxiliary variable x. Figure 1 provides a summary of the simulation setup we consider using a Directed Acyclic Graph (DAG). In Figure 1, Y and R are

Figure 1: A DAG for a setup where PMAR holds but SMAR does not hold. Variable U is latent in the sense that it is never observed.



conditionally independent given X, but they are not conditionally independent given X and I.

In order to make SMAR hold, we may seek to include design information in $\boldsymbol{x}$ to ensure that condition 1 of Lemma 1 holds. In this paper, we suppose that the only additional design information that can be used for this purpose consists of the design weights π_i^{-1} for sample units. We could include these weights in $\boldsymbol{x}$ but this still does not ensure SMAR as the example in Table 1 illustrates. Let $\pi_i^{-1} = 25$ when $(y_i, R_i) = (0, 0)$ or $(1, 1)$ and $\pi_i^{-1} = 6.25$ when $(y_i, R_i) = (0, 1)$ or $(1, 0)$, where these values have been obtained simply by inverting the proportions with $y = 1$ in the table. Then, we find that $P(y = 1 \mid R, I = 1, \pi_i^{-1} = 25)$ takes the value 0 if $R = 0$ and 1 if $R = 1$ and so SMAR does not hold even if we condition on π_i^{-1}.

Our goal now is to construct imputed values for which the imputed estimator in (1) is approximately unbiased under PMAR. Imputing from $f(y_i \mid \boldsymbol{x}_i, I_i = 1, R_i = 1)$, as before, will generally lead to bias if SMAR does not hold. For example, using this approach with the example in Table 1 will lead to only 20% of imputed values taking the value 1, whereas we would need this percentage to be 80% for the imputed estimator of the population proportion with $y = 1$ to be unbiased.

Imputing from $f(y_i \mid \boldsymbol{x}_i, I_i = 1, R_i = 0)$ and ensuring condition (2) does not seem feasible when SMAR fails, certainly not by imputing from a fitted model of $f(y_i \mid \boldsymbol{x}_i, I_i = 1, R_i = 1)$. Instead, we consider the alternative condition that the imputed values y_i^* satisfy

$$E\left\{y_i - y_i^* \mid \boldsymbol{x}_i, R_i = 0\right\} = 0. \tag{6}$$

The following lemma shows that condition (6) also leads to an unbiased imputed estimator.

Lemma 2 *Under (6), the imputed estimator of the form (1) is unbiased for θ in the sense that $E(\hat{\theta}_I - \theta) = 0$.*

Proof. Since

$$\hat{\theta}_I - \hat{\theta}_n = \sum_{i \in A} \pi_i^{-1}(1 - R_i)\{y_i^* - y_i\},$$

we have

$$E(\hat{\theta}_I - \hat{\theta}_n \mid R_1, \cdots, R_N, \mathcal{F}_N) = \sum_{i \in U}(1 - R_i)\,(y_i^* - y_i)$$

where the expectation is taken with respect to the sampling design. By (6), we have

$$E\{\sum_{i \in U}(1 - R_i)\,(y_i^* - y_i)\} = 0 \tag{7}$$

which gives the required result. ■

Condition (6) may be achieved under PMAR by noting that then

$$E(y_i \mid \boldsymbol{x}_i, R_i = 0) = E(y_i \mid \boldsymbol{x}_i, R_i = 1)$$

and so we have only to estimate the distribution $f(y_i \mid \boldsymbol{x}_i, R_i = 1)$, which is equal to $f(y_i \mid \boldsymbol{x}_i)$ under PMAR. Specifying $f(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}; \beta)$ as a parametric regression model, we can estimate the parameter vector β under informative sampling by using the sampling weights $w_i = \pi_i^{-1}$ and solving

$$\sum_{i \in A} w_i R_i S(\beta; \boldsymbol{x}_i, y_i) = 0, \tag{8}$$

where $S(\beta; \boldsymbol{x}_i, y_i) = \partial \log f(y_i \mid \boldsymbol{x}_i; \beta)/\partial\beta$ (Pfeffermann, 1993; Fuller, 2009). Once $\hat{\beta}$ is computed from (8), the imputed values y_i^* are generated from $f(y_i \mid \boldsymbol{x}_i; \hat{\beta})$ and the resulting estimator is approximately unbiased. This approach is referred to as the Weighting Method.

Our second approach is to consider an augmented regression model for $f(y \mid \boldsymbol{x}, w)$, where the sampling weight $w_i = 1/\pi_i$ or some function of it enters now as an additional

explanatory variable. The basic rationale for this approach is that conditioning on w renders the sampling ignorable in the sense that $f(y \mid \boldsymbol{x}, w) = f(y \mid \boldsymbol{x}, w, I = 1)$ (Rubin, 1987). Thus, in principle, we could fit a model to sample observations under informative sampling without any need for sample weighting. Since we are only interested in prediction rather than model parameters directly it does not matter that our model has changed.

A problem, however, is that we only have observations on y for $R = 1$. We can still estimate the distribution $f(y \mid \boldsymbol{x}, w, R = 1)$ by fitting a parametric model $f(y \mid \boldsymbol{x}, w, R = 1; \gamma)$ to cases with $I = 1$ and $R = 1$ without any need for sample weighting. In this case, the imputed value y_i^* can be generated from $f(y_i \mid \boldsymbol{x}_i, w_i, R_i = 1; \hat{\gamma})$ and we refer to this as the *Augmented Model Method*.

The problem is that in order to achieve condition (6), we should like PMAR to hold for the augmented model, that is: $f(y \mid \boldsymbol{x}, w, R = 1) = f(y \mid \boldsymbol{x}, w, R = 0)$. But this does not follow necessarily from the PMAR assumption $f(y \,|\, x, R = 1) = f(y \,|\, x, R = 0)$. Consider, for example, the set-up in Table 1, with values of w_i as described earlier. Then $P(y = 1 \mid w = 25, R = 1) = 1$ and $P(y = 1 \mid w = 25, R = 0) = 0$ so that, although PMAR holds unconditionally, it does not conditional on w. We shall illustrate the potential bias of the Augmented Model Method in the simulation study.

3 Fractional Imputation

Under either of the methods in the previous section, a single value y_i^* is imputed for each unit in the sample where y_i is missing. Either approach can be extended naturally to a fractional imputation approach where m imputed values $y_{i1}^*, \cdots, y_{im}^*$ are generated, with a view to improving efficiency of estimation of θ and enabling the use of replication variance estimation.

A general approach is obtained by taking $y_{i1}^*, \cdots, y_{im}^*$ to be generated from an arbitrary proposal distribution $f_0(y \mid \boldsymbol{x})$. For a parametric model assumption, $f(y \mid \boldsymbol{x}; \beta)$, a natural choice for the proposal distribution under the Weighting Method is $f_0(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}; \hat{\beta})$, where $\hat{\beta}$ is the solution to (8).

An alternative is to use a nonparametric proposal distribution. To generate m

imputed values $y_{i1}^*, \cdots, y_{im}^*$ from a nonparametric $f_0(y \mid \boldsymbol{x})$, one can use the following systematic sampling algorithm:

1. Generate $u_1 \sim U(0, 1/m)$.
2. Compute $u_j = u_1 + (j-1)/m$ for $j = 2, \cdots, m$.
3. For $j = 1, \ldots, m$, choose
$$y_{ij}^* = F_0^{-1}(u_j \mid \boldsymbol{x}_i) \tag{9}$$
where $F_0(y \mid \boldsymbol{x})$ is the cumulative distribution function derived from $f_0(y \mid \boldsymbol{x})$.

This approach removes the effect of Monte Carlo sampling by using the m quantiles of the proposal distribution $f_0(y \mid \boldsymbol{x})$ for the imputed values. This reduces the imputation variance to order $1/m^2$, rather than order $1/m$. In practice, to remove the discontinuity points of F_0, we use an interpolation technique when computing $F_0(y \mid \boldsymbol{x})$. That is, we can express the interpolated CDF $\tilde{F}_0(y \mid \boldsymbol{x})$ as,

$$\tilde{F}_0(y \mid \boldsymbol{x}) = F_0(y_{(i)} \mid \boldsymbol{x}) + (y - y_{(i)}) \frac{F_0(y_{(i+1)} \mid \boldsymbol{x}) - F_0(y_{(i)} \mid \boldsymbol{x})}{y_{(i+1)} - y_{(i)}} \quad \text{if } y_{(i)} \le y < y_{(i+1)},$$

where $y_{(i)}$ is the $i-th$ order statistic of $\{y_i : R_i = 1, I_i = 1\}$.

The fractional weight associated with y_{ij}^* is computed as

$$w_{ij}^* = \frac{f(y_{ij}^* \mid \boldsymbol{x}_i; \hat{\beta}) / f_0(y_{ij}^* \mid \boldsymbol{x}_i)}{\sum_{k=1}^m f(y_{ik}^* \mid \boldsymbol{x}_i; \hat{\beta}) / f_0(y_{ik}^* \mid \boldsymbol{x}_i)}. \tag{10}$$

Note that the fractional weight reduces to $w_{ij}^* = 1/m$ when $f_0(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}; \hat{\beta})$.

When m is small, the fractional weights can be further modified in the calibration step. The proposed calibration equation for improving the fractional weights in this case is

$$\sum_{i \in A} \sum_{j=1}^m w_i (1 - R_i) w_{ij}^* S(\hat{\beta}; \boldsymbol{x}_i, y_{ij}^*) = 0, \tag{11}$$

and $\sum_{j=1}^m w_{ij}^* = 1$ for each i with $R_i = 0$, where $\hat{\beta}$ is computed from (8). The calibration condition (11) guarantees that the imputed score equation leads to the same $\hat{\beta}$ (Kim and Shao, 2014, pg. 86-87). Then the fractionally imputed estimator of $\theta = \sum_{i=1}^N y_i$ is obtained by

$$\hat{\theta}_{FI} = \sum_{i \in A} w_i \left\{ R_i y_i + (1 - R_i) \sum_{j=1}^m w_{ij}^* y_{ij}^* \right\}.$$

We now consider variance estimation for the fractionally imputed estimator using a replication method. Replication variance estimation is very popular in practice. See Chapter 4 of Fuller (2009) for a comprehensive overview of the replication method for variance estimation. Let $\{w_i^{(k)} \mid i \in A\}$ be the k-th set of replication weights such that

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2 \tag{12}$$

is consistent for the variance of $\hat{\theta} = \sum_{i \in A} w_i y_i$, where L is the replication size, c_k is the k-th replication factor that depends on the replication method and the sampling mechanism (Fuller, 2009, Ch. 4), and $\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i$.

To apply the replication method to fractional imputation, we follow the approach of Kim and Shao (2014, pg. 91). First, apply the replication weights to compute $\hat{\beta}^{(k)}$ in (8). This is used to compute the replication fractional weights

$$w_{ij}^{*(k)} = \frac{f(y_{ij}^* \mid \boldsymbol{x}_i; \hat{\beta}^{(k)}) / f_0(y_{ij}^* \mid \boldsymbol{x}_i)}{\sum_{l=1}^{m} f(y_{il}^* \mid \boldsymbol{x}_i; \hat{\beta}^{(k)}) / f_0(y_{il}^* \mid \boldsymbol{x}_i)}$$

but the same imputed values y_{ij}^* are used for each replicate k. The following calibration equation

$$\sum_{i \in A} \sum_{j=1}^{m} w_i^{(k)} (1 - R_i) w_{ij}^{*(k)} S(\hat{\beta}^{(k)}; \boldsymbol{x}_i, y_{ij}^*) = 0$$

with $\sum_{j=1}^{m} w_{ij}^{*(k)} = 1$ is then used to obtain the final replicate fractional weights, as before. Once the replicated fractional weights are computed, then

$$\hat{\theta}_{FI}^{(k)} = \sum_{i \in A} w_i^{(k)} \left\{ R_i y_i + (1 - R_i) \sum_{j=1}^{m} w_{ij}^{*(k)} y_{ij}^* \right\}$$

can be used to compute the replication variance estimator

$$\hat{V}_{rep}(\hat{\theta}_{FI}) = \sum_{k=1}^{L} c_k (\hat{\theta}_{FI}^{(k)} - \hat{\theta}_{FI})^2. \tag{13}$$

The replication method is very useful for multipurpose estimation. For example, if another parameter of interest is $\phi = Pr(Y < 3)$, then the FI estimator of ϕ is computed by

$$\hat{\phi}_{FI} = \sum_{i \in A} w_i \left\{ R_i I(y_i < 3) + (1 - R_i) \sum_{j=1}^{m} w_{ij}^* I(y_{ij}^* < 3) \right\}$$

and its replication variance estimator is computed by

$$\hat{V}_{rep}(\hat{\phi}_{FI}) = \sum_{k=1}^{L} c_k(\hat{\phi}_{FI}^{(k)} - \hat{\phi}_{FI})^2,$$

where

$$\hat{\phi}_{FI}^{(k)} = \sum_{i \in A} w_i^{(k)} \left\{ R_i I(y_i < 3) + (1 - R_i) \sum_{j=1}^{m} w_{ij}^{*(k)} I(y_{ij}^* < 3) \right\}.$$

Remark 1 *It appears to be much harder to handle informative sampling using multiple imputation (MI). In MI, the point estimator* $\hat{\theta}_{MI} = m^{-1} \sum_{j=1}^{m} \hat{\theta}_{Ij}$ *is essentially the same as* $\hat{\theta}_{FI}$ *with* $w_{ij}^* = 1/m$*, since* $\hat{\theta}_{Ij}$ *is defined as* $\hat{\theta}_I$ *for the* j*-th imputed data set. The MI variance estimator is*

$$\hat{V}(\hat{\theta}_{MI}) = U_m + \left(1 + \frac{1}{m}\right) B_m, \tag{14}$$

where $U_m = m^{-1} \sum_{j=1}^{m} \hat{V}_{Ij}$*,* $B_m = (m-1)^{-1} \sum_{j=1}^{m} (\hat{\theta}_I^{(j)} - \hat{\theta}_{MI})^2$*, and* $\hat{V}_{Ij}$ *is the variance estimator, such as* $\hat{V}_{rep}$ *in (12), using the* j*-th imputed data set.*

To achieve consistency, it is usual to require that the imputation method obeys

$$E\left(U_m\right) = V\left(\hat{\theta}_n\right), \tag{15}$$

$$E\left(B_m\right) = V\left(\hat{\theta}_{MI} - \hat{\theta}_n\right) \tag{16}$$

and

$$Cov\left\{\hat{\theta}_n, \hat{\theta}_{MI} - \hat{\theta}_n\right\} = 0, \tag{17}$$

where $\hat{\theta}_n$ *is the full sample estimator that would be obtained if no data were missing. An approach which at least leads to an approximately unbiased MI point estimator is to use the Augmented model method, described in section 2, to generate the imputed values, that is to generate them from* $f(y \mid \boldsymbol{x}, w)$*. However, even if* $f(y \mid \boldsymbol{x}, w)$ *is correctly specified and the MI point estimator is consistent, the MI variance estimator is not necessarily consistent because conditions (15)-(17) do not hold under informative sampling.*

4 Simulation study

We compare the alternative imputation procedures and corresponding variance estimators through simulation, focusing on the situation in which PMAR holds but SMAR does not. The super-population model for the variable of interest y_i is

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{18}$$

where $e_i \sim \mathrm{N}(0, \sigma_e^2)$, $\beta_0 = -1.5$, $\beta_1 = 0.5$, $\sigma_e^2 = 1.04$, and $x_i \sim \mathrm{N}(2, 1)$. The response indicator R_i satisfies, $R_i \sim \text{Bernoulli}(\phi_i)$, where

$$\text{logit}(\phi_i) = -1 + 0.5x_i + 0.5u_i, \tag{19}$$

$u_i \sim \mathrm{N}(2, 1)$, and u_i is independent of x_i and e_i. The sampling design is Poisson sampling with sample membership indicator $I_i \sim \text{Bernoulli}(\pi_i)$, where

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 u_i + \alpha_2 y_i, \tag{20}$$

$\alpha_0 = -3$, $\alpha_1 = -1/3$, and $\alpha_2 = 0.1$. The generated finite populations in the simulation are of size $N = 50{,}000$. The selection probabilities are such that the median realized sample size is $\bar{n} = 1257$, and the response probabilities are such that the median number of respondents in a selected sample is $\bar{n}_r = 862$.

It is supposed that no design variables which directly determine π_i are observed. Instead, expression (20) captures the indirect dependence of π_i on y_i and on an unobserved variable u_i which is also associated with the response propensity ϕ_i. This implies a three-way association between y, I and R given x, as discussed in Section 2.

The following four estimation procedures are considered:

1. Procedure 1 (OLS, FI) is ordinary least squares (OLS) ignoring informative sampling; that is, it is a version of fractional imputation without using sampling weights to compute $\hat{\beta}$ in (11). The imputed value $y_{ij}^* \sim \mathrm{N}(\hat{\mu}_{i,ols}, \hat{\sigma}_{ols}^2)$, where $\hat{\mu}_{i,ols} = \hat{\beta}_{0,ols} + \hat{\beta}_{1,ols} x_i$, and $(\hat{\beta}_{0,ols}, \hat{\beta}_{1,ols}, \hat{\sigma}_{ols}^2)$ is the vector of OLS estimates of the parameters of (18).

2. Procedure 2 (WLS, FI) is the proposed fractional imputation (FI) procedure using the approach termed the Weighting Method in Section 2. The imputed value $y_{ij}^* \sim N(\hat{\mu}_{i,wls}, \hat{\sigma}_{wls}^2)$, where $\hat{\mu}_{i,wls} = \hat{\beta}_{0,wls} + \hat{\beta}_{1,wls} x_i$, $\hat{\boldsymbol{\beta}}_{wls} = (\hat{\beta}_{0,wls}, \hat{\beta}_{1,wls}, \hat{\sigma}_{wls}^2)$ satisfies $\boldsymbol{S}_w(\hat{\boldsymbol{\beta}}_{wls}) = \mathbf{0}$, $\boldsymbol{S}_w(\boldsymbol{\beta}) = \sum_{i=1}^N \pi_i^{-1} R_i I_i \boldsymbol{S}_i(\boldsymbol{\beta})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \sigma_e^2)$, and $\boldsymbol{S}_i(\boldsymbol{\beta})$ is the contribution from unit i to the score function corresponding to (18).

3. Procedure 3 (AUG, FI) is fractional imputation procedure using the approach termed the Augmented Model Method in Section 2. The extended model underlying Procedure 3 is

$$y_i = \boldsymbol{z}_i' \boldsymbol{\beta}_{aug} + e_i, \tag{21}$$

where $\boldsymbol{z}_i = (1, x_i, \text{logit}(\pi_i))$, and $e_i \sim \text{N}(0, \sigma_{e,aug}^2)$. The imputed value $y_{ij}^* \sim \text{N}(\boldsymbol{z}_i' \hat{\boldsymbol{\beta}}_{aug}, \hat{\sigma}_{e,aug}^2)$, where $\hat{\boldsymbol{\beta}}_{aug}$ and $\hat{\sigma}_{e,aug}^2$ are OLS estimates of the parameters of the augmented model (21).

4. Procedure 4 (AUG, MI) is multiple imputation, assuming the augmented model (21). The multiple imputation procedure is implemented with the software JAGS (Plummer, 2003). The priors for regression coefficients are independent normal distributions with mean zero and variance 10^6, and the prior for σ is uniform on the interval $[0, 10^6]$.

Table 2 and Table 3 provide summaries of the distributions of one simulated population. The π_i range from approximately 0.01 to 0.15, with an average sampling rate of 0.03. Because u_i is independent of y_i given x_i, the PMAR assumption holds. However, the partial correlation between y_i and $\text{logit}(\phi_i)$ given x_i and $\text{logit}(\pi_i)$ is 1, indicating that SMAR is severely violated.

The Monte Carlo (MC) sample size in the simulation is B =5,000. The variance of the simulation represents the joint design-model variance. The parameters of interest, θ, for the simulation are the finite population parameters $E(Y)$, $P(Y \leq 2)$, B_0, and B_1, defined by $E(Y) = N^{-1} \sum_{i=1}^N y_i$, $P(Y \leq 2) = N^{-1} \sum_{i=1}^N I(y_i \leq 2)$, and $(B_0, B_1)' = (\boldsymbol{X}_N' \boldsymbol{X}_N)^{-1} \boldsymbol{X}_N' \boldsymbol{y}_N$, where $\boldsymbol{X}_N = ((1, x_1)', \ldots, (1, x_N)')'$ and $\boldsymbol{y}_N = (y_1, \ldots, y_N)'$. The variances in Table 4 are variances of deviations between estimators

and parameters of interest. Because the parameters are functions of the finite population, the parameters vary between MC samples. In Procedures 1-3, these parameters are estimated by solving

$$\sum_{i\in A} w_i \left\{ R_i U(\theta; x_i, y_i) + (1-R_i) \sum_{j=1}^{m} w_{ij}^* U(\theta; x_i, y_{ij}^*) \right\} = 0$$

for the relevant estimating function $U(\theta; x, y)$, where $w_i = 1/\pi_i$, y_{ij}^* is the j-th imputed value in fractional imputation, and $w_{ij}^* = 1/m$. In Procedure 4, estimates are obtained as described in Rubin (1987). We used $m = 100$ for all imputation methods.

Table 4 summarizes the properties of the four estimation procedures. Procedure 2, the FI procedure using weighted least squares (WLS, FI), leads to approximately unbiased estimators of all parameters considered. The bias based on Procedure 1 (OLS, FI) is approximately two orders of magnitude larger than the bias based on Procedure 2 (WLS, FI) for $E(Y)$ and $P(Y \le 2)$. Procedure 1 is biased because SMAR is violated in the simulation; that is, $f(y \,|\, x, I = 1, R) \neq f(y \,|\, x, I = 1)$. The estimators based on the augmented model, procedures 3 (FI) and 4 (MI), are also biased because PMAR does not hold for the augmented model in this setup. In particular, $f(y \,|\, x, w, R = 1) \neq f(y \,|\, x, w)$. To see that $f(y \,|\, x, w, R = 1) \neq f(y \,|\, x, w)$, note that the covariance matrix of $(y, \text{logit}(\phi))$ given x and $\text{logit}(\pi)$ is

$$C\{(y, \text{logit}(\phi)) \,|\, x, \text{logit}(\pi)\} = \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & 0.25\sigma_z^2 \end{pmatrix} - \frac{1}{\alpha_1^2\sigma_z^2 + \alpha_2^2\sigma_e^2} \begin{pmatrix} \alpha_2^2\sigma_e^4 & \alpha_2\alpha_1\sigma_e^2\sigma_z^2 \\ \alpha_2\alpha_1\sigma_e^2\sigma_z^2 & \alpha_1^2\sigma_z^4 \end{pmatrix}, \quad (22)$$

which has non-zero off-diagonal elements.

Replication variance estimators are computed for Procedures 2-4. The replicate weights for Poisson sampling are computed by

$$\begin{aligned} w_i^{(k)} &= w_i + (1-\pi_i)^{0.5} w_i - (1-\pi_i)^{0.5} w_i^2, \quad i = k \\ &= w_i - (1-\pi_k)^{0.5} w_k w_i, \quad i \neq k, \end{aligned} \quad (23)$$

for $k = 1, \ldots, n$, where n is the realized sample size. The procedure defined in (13) is used to estimate the variance of the FI estimators, and the MI variance estimator defined in (14) is used for Procedure 4.

Table 5 contains the MC means of the variance estimators and the MC variances of the estimators in the simulation. The column "Ratio" in Table 5 is the ratio of the MC mean of the variance estimator to the MC variance of the corresponding estimator, and the column "t-stat" is an approximate $t-$test of the null hypothesis that $E[\hat{V}(\hat{\theta})] - V\{\hat{\theta}\} = 0$, where $\hat{V}(\hat{\theta})$ is the estimator of the variance of estimator $\hat{\theta}$ of θ, and the form of the test statistic is defined in Kim (2004). A $t-$statistic larger than 1.96 in absolute value suggests that the difference between $E[\hat{V}(\hat{\theta})]$ and $V\{\hat{\theta}\}$ exceeds the effect of MC variability. The variance estimators for Procedures 2 and 3 are approximately unbiased for the variances of the estimators of all the parameters considered. As explained by Yang and Kim (2016), the MI estimator of the variance of the estimator of the CDF has a large positive bias.

For the simulation model defined by (18-20), the conditional correlation between y_i and $\text{logit}(\phi_i)$ given x_i and $\text{logit}(\pi_i)$ is 1. To analyze a range of correlations, we consider a generalization of the model (18-20), where $\pi_i = cp_i$,

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 u_i + \alpha_2 y_i + \eta_i, \tag{24}$$

$$\text{logit}(\phi_i) = \gamma_0 + \gamma_1 x_i + \gamma_2 u_i + \delta_i, \tag{25}$$

$c \in [0, 1]$, $\delta_i \sim \text{N}(0, \sigma_\delta^2)$, $\eta_i \sim \text{N}(0, \sigma_\eta^2)$, $u_i \sim \text{N}(0, 1.25)$, and $x_i \sim \text{N}(2, 1.25)$. The role of c is to control the magnitude of π_i, while permitting flexible choices for α_1 and α_2 and avoiding extreme negative α_0. The critical components of (24) and (25) are the additional error terms δ_i and η_i, which allow the conditional correlation between y_i and $\text{logit}(\phi_i)$ given x_i and $\text{logit}(p_i)$ to be less than 1. For the simulations discussed below, we set $c = 0.05$ and $\sigma_\eta = \sigma_\delta = 0.2$. We consider four parameter configurations that generate a full factorial defined by high and low levels of $|Cor(\pi_i, \phi_i)|$ and $|Cor(y_i, \text{logit}(\phi_i) \,|\, x_i, \text{logit}(p_i))|$. Table 6 gives the parameter configurations and corresponding sample correlations and partial correlations. The setting denoted (A, B) in Table 6 indicates that $Cor(\pi_i, \phi_i)$ is at level A, and $Cor(y_i, \text{logit}(\phi_i) \,|\, x_i, \text{logit}(p_i))$ is at level B, where $(A, B) \in \{\text{High}, \text{Low}\} \times \{\text{High}, \text{Low}\}$.

To conserve space, we summarize the MC properties of the estimators of the mean of y_i and present tabular output in online supplementary material. The (High,

High) setting is similar to the first simulation model defined by (18)-(19) in that $Cor(\pi_i, \phi_i)$ and $Cor(y_i, \text{logit}(\phi_i) \,|\, x_i, \text{logit}(p_i))$ are both large in absolute value. As expected, the Weighting Method (WLS, FI) dominates the other procedures in terms of MC MSE for the (High, High) setting. The OLS estimator is biased for the (High, High) configuration because SMAR is strongly violated, and the augmented model procedures (both MI and FI) are biased because PMAR does not hold for the augmented model. The (High, Low) parameter configuration is informative because the partial correlation between y_i and $\text{logit}(\phi_i)$ given x_i and $\text{logit}(p_i)$ is 0. For the (High, Low) setting, the SMAR assumption holds, although the correlation between π_i and ϕ_i is relatively large. Because SMAR holds, the OLS estimator is more efficient than the other estimators for the (High, Low) setting. This simulation configuration demonstrates that although the correlation between π_i and ϕ_i is related to whether or not the design is informative for the response model, this correlation is less relevant to a study of the SMAR assumption. To explain why, consider $\alpha_2 = 0$. For $\alpha_2 = 0$, increasing α_1 can increase the conditional correlation between $\text{logit}(p_i)$ and $\text{logit}(\phi_i)$ given x_i, although SMAR is satisfied for any α_1. The (Low, High) parameter configuration demonstrates the impact of the partial correlation between y_i and $\text{logit}(\phi_i)$ given x_i and $\text{logit}(p_i)$. Although the correlation between ϕ_i and π_i is low, the estimator based on the Weighting Method (WLS, FI) has smaller MC MSE than the alternative estimators for the (Low, High) parameter configuration. For the (Low, Low) parameter set, the procedures that incorporate the weights are more efficient than the OLS procedure. The OLS procedure is biased because SMAR is violated, though weakly. The Augmented Model Methods (both MI and FI) have MC MSEs modestly smaller than that of the Weighting Method for the (Low, Low) parameter configuration because the MC variances of the augmented model estimators are smaller than the MC variance of the WLS FI estimator, and the variance dominates the bias for the (Low, Low) simulation configuration. However, the absolute MC biases of the augmented model estimators are approximately one order of magnitude larger than the absolute MC bias of the WLS FI estimator for the (Low, Low) setting.

5 Comparison of Fractional Imputation Estimators for the Conservation Effects Assessment Project

The Conservation Effects Assessment Project (CEAP) is a survey that collects data intended to quantify different types of water and wind erosion. The sample design for CEAP is a two-phase sample. The first phase is based on the National Resources Inventory (NRI), a larger survey that monitors characteristics related to agriculture and natural resources, such as land cover, land use, and erosion, on non-federal US land. The CEAP sample is a subset of locations (longitude, latitude) classified as cultivated cropland in the NRI. Because typical sampling rates are less than 5%, approximating the CEAP sample as a with replacement sample is considered reasonable. Berg and Yu (2015) provides further detail on the sample design for CEAP and explains how first and second order inclusion probabilities are calculated.

The unit of analysis in CEAP is the crop field containing the sampled location. The farmer who operates the selected crop field is asked to complete an extensive questionnaire that requests detailed information on crops planted and conservation practices employed. Nonresponse arises in CEAP when farmers refuse to complete the questionnaire.

The data from the farmer interview survey, in conjunction with NRI data and administrative information on soil characteristics, are input to a physical process model called the Agricultural Policy Environmental Extender (APEX) that outputs several measures of erosion. One component of the APEX model is the Revised Universal Soil Loss Equation (RUSLE2). The RUSLE2 computer model transforms the collected data to a measure of a particular type of water erosion called sheet and rill erosion. The RUSLE2 model for sheet and rill erosion is an advancement of a traditional approximation called the Universal Soil Loss Equation (USLE). USLE is a product of five indexes related to crop managements, conservation practices, rainfall, soil erodibility, and slope length and steepness. RUSLE2 extends USLE by incorporating daily weather and more detailed information on cropping systems, for example. While RUSLE2 is only available for the respondents to the CEAP survey,

the NRI provides USLE for the full CEAP sample.

This analysis explores imputation of RUSLE2 using USLE as a covariate for a subset of the data from a national CEAP survey conducted over the period 2003-2006. We restrict to data collected during 2003-2005 because the sample design for 2006 differs from that used for the previous three years, and information to calculate selection probabilities for the 2006 sample is unavailable. We consider seven states that comprise the majority of the Corn Belt region, one of ten CEAP Production Regions defined for purposes of sampling and estimation. Table 7 gives the sample sizes and number of respondents for the seven Corn Belt states. The response rates range from 60% to 70% in these seven states.

Because the erosion measurements have skewed distributions, the imputation model is applied after transforming both RUSLE2 and USLE. Visual inspection and experimentation suggest a transformation of a power of 0.2. In the left panel of Figure 2, $\text{RUSLE2}^{0.2}$ is plotted against $\text{USLE}^{0.2}$ for the complete cases in the Indiana data. The right plot of Figure 2 contains the corresponding normal quantile-quantile plot of the residuals from the ordinary least squares regression of $\text{RUSLE2}^{0.2}$ on $\text{USLE}^{0.2}$ for the Indiana data. The linearity of the plots in Figure 2 suggest that the assumption of a linear relationship between $\text{RUSLE2}^{0.2}$ and $\text{USLE}^{0.2}$ with normally distributed errors may be reasonable.

5.1 Evaluating the Need for Weights in Estimating the Imputation Model for CEAP

Several aspects of the sample design and potential nonresponse mechanisms may cause one to question an assumption of SMAR. For example, one of the essential components of the design of the second-phase sample that defines the 2003-2005 CEAP sample is a stratification of the locations in the NRI first phase sample. One of the strata used for the second phase sample contains NRI sampling units with characteristics of high erosion. As a result of this definition of the stratification, a nontrivial relationship between the selection probabilities and the measures of sheet and rill erosion (USLE and RUSLE2) may be expected. Other features of the NRI

sample design are more complex and involve detailed stratification. We view the sample weights as the best available information on the NRI sample design to use in modeling.

Because the OLS and WLS estimators of the parameters of the super-population model are highly correlated, the standard error of either estimator is a poor indication of whether the means of the OLS and WLS estimators differ. To formally evaluate the need to use the weights to estimate the imputation model in the CEAP survey, we consider two test procedures. The first test procedure is based on the distribution of the difference between the weighted and unweighted estimators of regression coefficients. The second uses an expanded model motivated by the simulation model of Section 4.

To define the procedures, we formalize the imputation model for CEAP. In the superpopulation, assume

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{26}$$

where $e_i \sim \mathrm{N}(0, \sigma_e^2)$, $y_i = \mathrm{RUSLE2}^{0.2}$, and $x_i = \mathrm{USLE}^{0.2}$. Assume the PMAR condition (5) holds.

The null hypothesis for the first test procedure is

$$H_0 : E[\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_{ols}] = \mathbf{0}, \tag{27}$$

where $\hat{\boldsymbol{\beta}}_w = \boldsymbol{H}_{wls}^{-1} \sum_{i=1}^n \boldsymbol{x}_i R_i \pi_i^{-1} y_i$, $\hat{\boldsymbol{\beta}}_{ols} = \boldsymbol{H}_{ols}^{-1} \sum_{i=1}^n \boldsymbol{x}_i R_i y_i$, $\boldsymbol{H}_{wls} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' R_i \pi_i^{-1}$, $\boldsymbol{H}_{ols} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' R_i$, $\boldsymbol{x}_i = (1, x_i)'$, and the sample size is n. Define the test statistic

$$Q_1 = (\hat{\boldsymbol{\beta}}_{ols} - \hat{\boldsymbol{\beta}}_{wls})' \hat{\boldsymbol{V}}^{-1} (\hat{\boldsymbol{\beta}}_{ols} - \hat{\boldsymbol{\beta}}_{wls}), \tag{28}$$

where $\hat{\boldsymbol{V}}$ is a design consistent estimator of the variance of $\sum_{i=1}^n R_i \pi_i^{-1} (\hat{\boldsymbol{\xi}}_{i,ols} - \hat{\boldsymbol{\xi}}_{i,wls})$, $\hat{\boldsymbol{\xi}}_{i,ols} = \boldsymbol{H}_{ols}^{-1} \pi_i \boldsymbol{d}_i(\hat{\boldsymbol{\beta}}_{ols})$, $\hat{\boldsymbol{\xi}}_{i,wls} = \boldsymbol{H}_{wls}^{-1} \boldsymbol{d}_i(\hat{\boldsymbol{\beta}}_{ols})$, and $\boldsymbol{d}_i(\hat{\boldsymbol{\beta}}_{ols}) = (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{ols}) \boldsymbol{x}_i$. Replication procedures may be used instead to obtain $\hat{\boldsymbol{V}}$. Under the null hypothesis (27), Q_1 has an approximate chi-squared distribution with 2 degrees of freedom.

For the second test procedure, we consider the extended model

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 \mathrm{logit}(\pi_i) + \theta_3 \mathrm{logit}(\pi_i) x_i + b_i, \tag{29}$$

where $b_i \sim \mathrm{N}(0, \sigma_b^2)$. The null hypothesis for the second test procedure is

$$H_0 : E[\hat{\theta}_2] = E[\hat{\theta}_3] = 0, \tag{30}$$

where $\hat{\theta}_2$ and $\hat{\theta}_3$ are the OLS estimators of θ_2 and θ_3. To define the test statistic for (30), let $\hat{\boldsymbol{\theta}}$ be the OLS estimator of $(\theta_0, \theta_1, \theta_2, \theta_3)'$, and let $\hat{\boldsymbol{V}}_a$ be an estimator of the variance of $\hat{\boldsymbol{\theta}}$. One choice of $\hat{\boldsymbol{V}}_a$ is a design consistent estimator of the variance of $\sum_{i=1}^n R_i \pi_i^{-1} \boldsymbol{H}_a^{-1} \boldsymbol{d}_{i,a}(\hat{\boldsymbol{\theta}})$, where $\boldsymbol{d}_{i,a}(\hat{\boldsymbol{\theta}}) = \pi_i(y_i - \boldsymbol{z}_i'\hat{\boldsymbol{\theta}})\boldsymbol{z}_i$, and $\boldsymbol{z}_i = (1, x_i, \mathrm{logit}(\pi_i), \mathrm{logit}(\pi_i)x_i)'$. Replication procedures may be used instead to obtain $\hat{\boldsymbol{V}}_a$. Define the test statistic

$$Q_2 = \hat{\boldsymbol{\theta}}_{23}' \hat{\boldsymbol{V}}_{23}^{-1} \hat{\boldsymbol{\theta}}_{23}, \tag{31}$$

where $\hat{\boldsymbol{\theta}}_{23}$ is the the OLS estimator of $(\theta_2, \theta_3)'$, and $\hat{\boldsymbol{V}}_{23}$ is the sub-matrix of $\hat{\boldsymbol{V}}_a$ corresponding to $(\theta_2, \theta_3)'$. Under the null hypothesis (30), Q_2 has an approximate chi-squared distribution with 2 degrees of freedom.

The simulations with the (High, Low) and (High, High) parameter settings defined in Table 6 vet the test procedures defined by (28) and (31). For the (High, Low) setting, the null hypotheses (27) and (30) hold, and the statistics (28) and (31), respectively, exceed the 95th percentile of a chi-squared distribution with two degrees of freedom in 4.1% and 5.2% of the 5,000 MC samples. For the (High, High) simulation, the test procedures reject at the nominal 5% level for all 5,000 MC samples.

Table 8 contains several statistics related to the use of weights in estimating the imputation model for the CEAP data. The columns $p(Q_1)$ and $p(Q_2)$, respectively, are the $p-$values corresponding to Q_1 and Q_2, using a chi-squared distribution with 2 degrees of freedom as a reference distribution. Both tests reject the respective null hypotheses at the 5% level for the same states. The columns $Cor(v_i, \pi_i)$ give the correlation between variable v_i and π_i for $v_i = y_i, r_i(\hat{\boldsymbol{\beta}}_{ols})$, and $r_i(\hat{\boldsymbol{\beta}}_{ols})x_i$, where $r_i(\hat{\boldsymbol{\beta}}_{ols}) = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{ols})$. As expected, $Cor(r_i(\hat{\boldsymbol{\beta}}_{ols}), \pi_i)$ and $Cor(r_i(\hat{\boldsymbol{\beta}}_{ols})x_i, \pi_i)$ are relatively small for the states where neither null hypothesis is rejected. That the null hypotheses are rejected at the 5% level for three states (IL, IN, WI) and that the

$p-$values are close to 5% for OH provide support for using the weights to estimate the imputation model.

An examination of the estimates of the coefficients in the expanded model (29) is interesting. Table 9 contains estimates of $(\theta_0, \theta_1, \theta_2, \theta_3)'$ and corresponding $t-$statistics. For IL and WI, the $t-$statistics indicate that the null hypothesis (30) is rejected because the coefficient associated with $\text{logit}(\pi_i)x_i$ differs significantly from 0, rather because the coefficient associated with $\text{logit}(\pi_i)$ differs significantly from 0. This indicates that an expanded model with only $\text{logit}(\pi_i)$ as the additional explanatory variable may be inadequate.

5.2 Estimates of Mean RUSLE2

The parameter of interest is the mean RUSLE2 soil loss for the state defined, $\theta = E[y_i^5]$. The number of imputed values $J = 100$. We compare estimates of mean RUSLE2 based on the three FI procedures used for the simulation: 1 (OLS – least squares, ignoring informative sampling), 2 (WLS – Weighting Method), and 3 (Augmented Model Method, with $\text{logit}(\pi_i)$ as the explanatory variable and ordinary least squares estimates).

Table 10 contains estimates of mean RUSLE2 and corresponding estimated standard errors based on the three FI procedures. Taylor linearization is used to calculate the standard errors (Kim and Shao, 2014, pg. 69). The differences between the estimates based on weighted least squares and the estimates based on ordinary least squares are larger for IL, IN, and WI than for IA, MI, MN, and OH. This pattern is consistent with the test statistics in Table 8 and the observation that IL, IN, and WI are the states for which the correlations between the residuals of the ordinary least squares regressions of y_i on x_i and the selection probabilities are relatively large. With the exception of IA and MN, the augmented model estimators lie between the corresponding OLS and WLS estimators.

The $t-$statistics provide only marginal support for the augmented model. The $t-$statistics for θ_2 are less than 2 for all states. The $t-$statistic for θ_3 for Indiana is less than 1.96. Because RUSLE2 and USLE measure the same kind of erosion, the

theoretical relationship between RUSLE2 and USLE does not intuitively support the the expanded model (29).

With the exception of MI, the weighted least squares procedure produces smaller estimates of mean RUSLE2 than ordinary least squares. Figure 3 illustrates how this difference in estimated means relates to the estimated regression coefficients and imputed values. The black points in Figure 3 are the pairs (x_i, y_i) for CEAP respondents. The red and green lines are the regression lines based on ordinary least squares and weighted least squares, respectively. The red points correspond to the pairs $(x_i, \tilde{y}_{i,ols})$, where $\tilde{y}_{i,ols} = J^{-1} \sum_{j=1}^{J} y_{ij,ols}$, and $y_{ij,ols}$ is the j^{th} imputed value for nonrespondent i based on ordinary least squares. The green points correspond to the pairs $(x_i, \tilde{y}_{i,wls})$, where $\tilde{y}_{i,wls} = J^{-1} \sum_{j=1}^{J} y_{ij,wls}$, and $y_{ij,wls}$ is the j^{th} imputed value for nonrespondent i based on weighted least squares. For IL, IN, and WI, the ordinary least squares estimate of the slope is larger than the weighted least squares estimate, resulting in larger imputed values for larger x_i.

6 Concluding Remarks

In this paper, we have considered imputation in a setting where missingness is ignorable in the population (PMAR) but not in the sample (SMAR). Such a circumstance might arise when the sample inclusion probabilities π_i are related not only to survey outcome variables y_i of interest but also to response probabilities ϕ_i, after conditioning on observed covariates. This covariance structure may arise via some shared dependence on an unobserved variable, such as u_i in the simulation study. In such a setting, we have observed that bias may arise not only for conventional imputation which ignores the sampling scheme but also for the augmented model approach which has been used for informative sampling, in which the imputation model is augmented to include the sampling weight. The empirical results demonstrate that procedures based on augmented models that incorporate the selection probabilities may lead to biased estimators if the assumption of population missing at random does not hold for the extended model. A current practice of multiple imputation under informative sampling, based on the augmented model approach, is still subject to this problem.

We have shown that such bias can be avoided by appropriately incorporating the sampling weights into the estimating equation for imputation model parameters. We accomplish this through fractional imputation and demonstrate how to obtain design consistent variance estimators for the imputation based estimators through replication procedures. We compare estimators of mean erosion based on the three fractional imputation methods using data from CEAP. Test procedures support the use of the weighted estimator for the CEAP data.

Supplementary Material

Please see the online supplement titled "Supplement to Imputation under informative sampling" for tabular output corresponding to the second set of simulations.

References

Berg, E.J. and Yu, C.L. (2015). Semiparametric quantile regression imputation for a complex survey. Unpublished Manuscript.

Carpenter, J. R. and Kenward, M.G. (2013). *Multiple Imputation and its Application*. Wiley, Chichester.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, **98**, 459–471.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion), *Journal of the Royal Statistical Society: Series B*, **41**, 1–31.

Deville, J.-C. and Särndal, C.-E.(1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381–394.

Fay, R. E. (1992). When are inferences from multiple imputation valid? In *Proceeding in Survey Research Method Section.,*pp. 227–32. Washington, DC: American Statistical Association.

Fuller, W.A. (2009). *Sampling Statistics*, Wiley, Chichester.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann and C.R. Rao, eds. *Sample Surveys: Design, Methods and Applications*, Elsevier, Amsterdam, 215–246.

Kim, J.K. (2004). Finite Sample Properties of Multiple Imputation Estimators. *Annals of Statistics*, 32, 766–783.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika*, **98**, 119–132.

Kim, J.K. and Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96**, 917–932.

Kim, J.K. and Shao, J. (2014). *Statistical Methods for Handling Incomplete Data*. CRC Press, Boca Raton.

Kott, P. S. (1995). A paradox of multiple imputation. In *Proceeding in Survey Research Method Section.,*pp. 380–383. Washington, DC: American Statistical Association.

Little, R.J.A. (2003). Bayesian methods for unit and item nonresponse. In Chambers, R.L. and Skinner, C. J. *Analysis of Survey Data*. Wiley, Chichester, 289–306.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *International Statistical Review*, **61**, 317–337.

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

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166–186.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian Graphical Models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Dis-*

tributed Statistical Computing (DSC 2003), March 20-22, Vienna, Austria. ISSN 1609-395X.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, **32**, 143-149.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion) *Journal of the American Statistical Association*, **91** 473–489.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Schenker, N., Raghunathan, T.E., Chiu, P-L., Makuc, D.M., Zhang G. and Cohen A.J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, **101** 924–933.

Seaman, S.R., White, I.R., Copas, A.J. and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics*, **68** 129–137.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by Missing at Random? *Statistical Science*, **28**, 257–268.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, **94**, 254–265.

Skinner, C., Holt, D. and Smith, T.M.F. eds. (1989). *Analysis of Complex Surveys*. Chichester, Wiley.

Yang, S. and Kim, J.K. (2016). A Note on Multiple Imputation for General-Purpose Estimation, *Biometrika*, In press (available at *doi: 10.1093/biomet/asv073*).

	$Y=0$		$Y=1$	
	$R=0$	$R=1$	$R=0$	$R=1$
$I=0$	240	210	210	240
$I=1$	10	40	40	10

Table 1: Frequency table illustrating how PMAR may hold but SMAR does not

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
x_i	-1.23	1.35	2.01	2.01	2.69	5.43
y_i	-4.90	-1.21	-0.51	-0.48	0.25	3.80
π_i	0.01	0.02	0.02	0.03	0.03	0.15
ϕ_i	0.15	0.63	0.73	0.71	0.81	0.97

Table 2: Summaries of distributions of variables in the simulation for one generated population.

	x_i	y_i	π_i	ϕ_i
x_i	1.000	0.441	0.143	0.697
y_i	0.441	1.000	0.318	0.305
π_i	0.143	0.318	1.000	-0.550
ϕ_i	0.697	0.305	-0.550	1.000

Table 3: Sample correlation matrix of variables in the simulation for one generated population.

Parameter	Procedure	MSE	Variance	Bias
$E(Y)$	1 (OLS, FI)	0.00241	0.00154	0.02952
	2 (WLS, FI)	0.00163	0.00163	0.00042
	3 (AUG, FI)	0.00318	0.00148	0.04119
	4 (AUG, MI)	0.00317	0.00148	0.04102
$P(Y \leq 2)$	1 (OLS, FI)	0.00033	0.00023	-0.00980
	2 (WLS, FI)	0.00024	0.00024	-0.00004
	3 (AUG, FI)	0.00042	0.00023	-0.01382
	4 (AUG, MI)	0.00042	0.00023	-0.01384
B_0	1 (OLS, FI)	0.00999	0.00755	0.04937
	2 (WLS, FI)	0.00804	0.00804	0.00129
	3 (AUG, FI)	0.01182	0.00724	0.06765
	4 (AUG, MI)	0.01177	0.00726	0.06715
B_1	1 (OLS, FI)	0.00141	0.00132	-0.00975
	2 (WLS, FI)	0.00139	0.00139	-0.00027
	3 (AUG, FI)	0.00145	0.00128	-0.01307
	4 (AUG, MI)	0.00144	0.00128	-0.01290

Table 4: Monte Carlo MSE, variance, and bias of estimation procedures.

Parameter	Procedure	Ratio	"t-stat"
$E[Y]$	2 (WLS, FI)	0.9920	-0.4063
$P(Y \leq 2)$	2 (WLS, FI)	0.9769	-1.1670
B_0	2 (WLS, FI)	0.9703	-1.4524
B_1	2 (WLS, FI)	0.9735	-1.2961
$E[Y]$	3 (AUG, FI)	0.9813	-0.9538
$P(Y \leq 2)$	3 (AUG, FI)	0.9997	-0.0150
B_0	3 (AUG, FI)	0.9629	-0.8185
B_1	3 (AUG, FI)	0.9787	-1.0413
$E[Y]$	4 (AUG, MI)	1.0017	0.0828
$P(Y \leq 2)$	4 (AUG, MI)	1.1789	9.0506
B_0	4 (AUG, MI)	1.0339	1.6818
B_1	4 (AUG, MI)	1.0398	1.9884

Table 5: Comparison of MC means of estimators of variances to MC variances of estimators based on Procedures 2-4. The column "Ratio" is the ratio of the MC mean of the variance estimator to the MC variance of the estimator. The column "t-stat" is an approximate t-test of the null hypothesis that the ratio of the mean of the variance estimator to the variance of the estimator is 1.

Setting	γ_2	α_1	α_2	$Cor(\pi_i, \phi_i)$	$Cor(y_i, \text{logit}(\phi_i) \mid x_i, \text{logit}(p_i))$
(High, High)	1	0.5	0.5	0.54	-0.86
(High, Low)	1	1	0	0.54	0
(Low, High)	1	1	-2.5	0.11	0.97
(Low, Low)	0.1	0.75	-0.15	0.07	0.10

Table 6: Parameter values and corresponding correlations for the model defined by (18), (24), and (25). For all four sets, $\beta_0 = -1.5$, $\beta_1 = 0.5$, $\gamma_0 = 0.5$, $\gamma_1 = 0.5$, and $\alpha_0 = -3.5$.

State	Sample Size	Number of Respondents
Illinois (IL)	1823	1275
Indiana (IN)	1151	751
Iowa (IA)	1492	1011
Michigan (MI)	935	585
Minnesota (MN)	1649	1008
Ohio (OH)	1053	698
Wisconsin (WI)	662	414

Table 7: Sample sizes and number of respondents for the seven Corn Belt states.

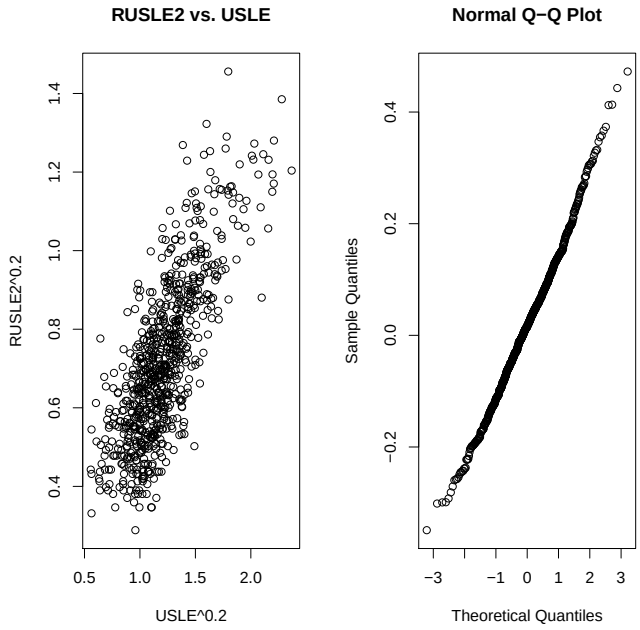


Figure 2: Illustration of relationships between RUSLE2 and USLE for IN. Left: $\text{RUSLE2}^{0.2}$ vs. $\text{USLE}^{0.2}$. Right: Residuals from the ordinary least squares regression of $\text{RUSLE2}^{0.2}$ on $\text{USLE}^{0.2}$ vs. the quantiles of a normal distribution.

ST	$Cor(y_i, \pi_i)$	$Cor(r_i(\hat{\boldsymbol{\beta}}_{ols}), \pi_i)$	$Cor(r_i(\hat{\boldsymbol{\beta}}_{ols})x_i, \pi_i)$	Q_1	$p(Q_1)$	Q_2	$p(Q_2)$
IL	0.242	0.126	0.131	19.643	0.000*	21.924	0.000
IN	0.219	0.205	0.200	32.839	0.000	38.917	0.000
IA	0.110	0.054	0.056	0.997	0.608	0.984	0.611
MI	0.170	-0.045	-0.063	2.436	0.296	3.522	0.172
MN	0.143	0.051	0.063	3.121	0.210	4.558	0.102
OH	0.085	0.044	0.031	5.653	0.059	4.974	0.083
WI	0.097	0.143	0.158	15.923	0.000	18.625	0.000

Table 8: Sample correlations, test statistics defined in (28) and (31), and corresponding $p-$values for the CEAP data. Here, $r_i(\hat{\boldsymbol{\beta}}_{ols}) = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{ols}$. A *0.000 means p-value<0.001.

	Estimates				t-statistics			
	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
IL	0.040	0.645	-0.032	0.040	0.332	6.717	-1.345	2.103
IN	0.025	0.726	-0.012	0.042	0.189	6.705	-0.433	1.850
IA	0.318	0.352	0.014	-0.006	1.734	2.368	0.398	-0.212
MI	0.404	0.306	0.045	-0.047	3.256	2.724	1.591	-1.781
MN	0.206	0.443	-0.022	0.027	2.401	5.766	-1.214	1.601
OH	0.450	0.332	0.050	-0.037	3.034	2.339	1.595	-1.214
WI	0.139	0.701	-0.060	0.080	0.727	4.534	-1.452	2.383

Table 9: Estimates and $t-$statistics for the parameters of the expanded model (29).

	OLS		WLS		AUG	
ST	Mean	SE	Mean	SE	Mean	SE
IL	0.336	0.009	0.327	0.009	0.332	0.009
IN	0.300	0.013	0.285	0.013	0.290	0.013
IA	0.345	0.011	0.339	0.011	0.343	0.011
MI	0.313	0.015	0.315	0.015	0.315	0.015
MN	0.166	0.004	0.166	0.004	0.164	0.004
OH	0.365	0.016	0.360	0.016	0.362	0.016
WI	0.516	0.024	0.490	0.024	0.500	0.024

Table 10: Comparison of estimates of mean RUSLE2 based on three FI procedures: 1 (OLS), 2 (WLS), and 3 (AUG). The three procedures are defined as in the simulation.

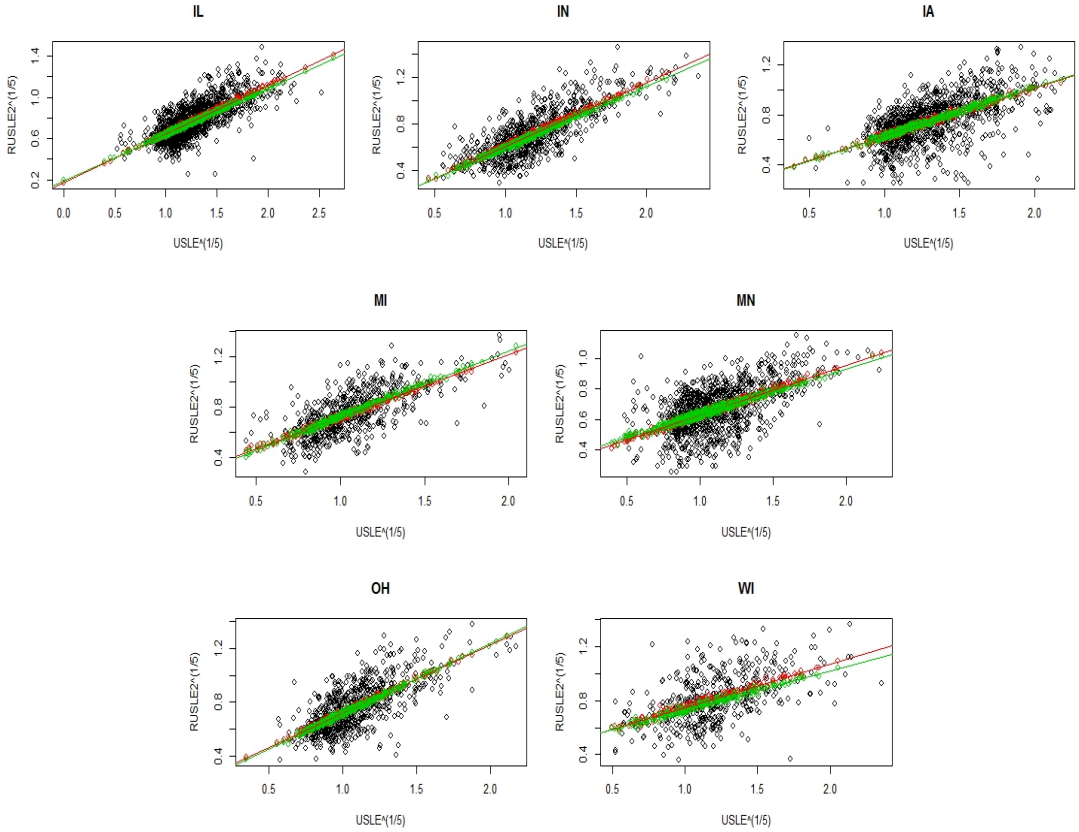


Figure 3: $\text{RUSLE2}^{0.2}$ vs. $\text{USLE}^{0.2}$ for Corn Belt states. Black = observed. Red = imputed values and regression line based on OLS. Green = imputed values and regression line based on WLS.