REPLICATION VARIANCE ESTIMATION FOR THE NATIONAL RESOURCES INVENTORY

J.D. Opsomer, W.A. Fuller and X. Li
Iowa State University, Ames, IA 50011, USA

1. Introduction

Replication methods are often used in variance estimation and are particularly appealing in situations where the design and/or the estimation procedure are complex. In this article, we focus on replication methods for two-phase regression estimators, and in particular, we consider the situation in which auxiliary information is available for both the population and the first phase of sampling. Two-phase regression estimators can be calibrated for the population information, for the phase one estimates or for both "levels" of information. For a discussion of different types of regression estimators in this context, see Estevao and Särndal (2002) and also Li and Opsomer (2003).

Jackknife variance estimators in two-phase sampling are derived by Rao and Sitter (1995) for ratio estimation and Sitter (1997) for regression estimation. These methods use the entire first-phase sample. However, if the number of elements in the phase one sample is large, these methods are computationally cumbersome. Fuller (1998) develops a variance estimator that makes it possible to compute the full two-phase variance using only the second-phase sample, by incorporating the phase one variability of the phase one calibration variables into the phase two replicates. Recently, Kim and Sitter (2003) discuss a different approach for avoiding the need for complete phase one replicates for several specific designs.

In this article, we describe a variance estimator closely related to that of Fuller (1998), and discuss its implementation for variance estimation of the National Resources Inventory (NRI). The remainder of the article is organized as follows. Section 2 reviews the sampling design and estimation for the NRI. In Section 3, we discuss our proposed variance estimator for two-phase regression estimators, and Section 4 gives simulation results. In Section 5, we conclude by describing the implementation to the NRI.

Department of Statistics, Iowa State University, Ames, IA 50011, USA; jopsomer@iastate.edu.

2. National Resources Inventory

The National Resources Inventory (NRI) is a statistical survey of land use and natural resource conditions and trends on U.S. non-federal lands. The NRI is conducted by the Natural Resources Conservation Service (NRCS) of the U.S. Department of Agriculture in cooperation with Iowa State University's Center for Survey Statistics and Methodology (CSSM).

The NRI was conducted every 5 years during the period 1982–1997. The 1997 NRI contains approximately 300,000 areal plots (or *segments*) distributed according to a spatially stratified design. While some data elements are collected for the segments, the majority of the data elements of interest are recorded at point locations within these segments. The total number of 1997 NRI points is approximately 800,000. Since 2000, the full panel structure of the NRI has been replaced by a two-phase sampling design, in which the 1997 NRI segments serve as a first phase, and each year a partially overlapping panel is selected through a stratified sampling design as a second phase. The annual second phase samples include approximately 42,000 "core" segments that are to be visited every year. An additional 30,000 segments are selected from the remaining 268,000 each year to form a supplemental sample. All points in all selected segments are part of the annual sample. See Nusser and Goebel (1997) and Fuller (2003b) for a more complete description of the NRI sampling design.

Each year, the estimation procedure combines information from several sources to produce a final data set composed of records containing information for the years 1982, 1987, 1992, 1997, 2000, and annually thereafter. First, the data collected at the segment level are converted to point data. For each of these new points (referred to as *pseudo-points*), a hierarchical hot-deck imputation procedure is used for filling in the unobserved point data elements, and an initial weight is assigned. The initial weights for both observed points and pseudo-points are adjusted during the estimation process using ratio and small area estimation. Control totals for surface area, federal land, and large water areas, derived from GIS databases, are maintained throughout the process. Finally, the weights are adjusted using iterative proportional scaling (raking) so that areas estimated for major broad coveruses for historical years in the current survey closely match those earlier estimates. For further details on the NRI estimation procedure, see Fuller (1999) and Fuller (2003a).

Until 1997, variance estimation for the NRI was based on the linearized approximation approach (e.g. Särndal et al. 1992, Ch. 5), and computed by an algorithm based on PC-CARP (Fuller et al. 1986). With the introduction of the annual NRI, this approach became impractical. In addition to the significant methodological difficulties in deriving estimators appropriate for capturing the complete sampling and estimation procedures described above, the resulting procedure would almost certainly require access to both the phase one and the phase two samples in order to compute the necessary variance components. This was considered undesirable from a practical standpoint, not only because of the obvious increase in storage requirements, but also because of the potential for confusion by the data analysts, who would be able to compute different estimates for the same quantity from both provided datasets. For these reasons, it was decided to investigate replicate variance estimation for the annual NRI surveys as an alternative to linearized-based estimation.

3. Replicate Variance Estimation

We begin by describing the two-phase regression estimator and a corresponding replicate variance estimator in a simple setting. Let s_1 and s_2 denote the first-phase and second-phase samples of size n_1 and n_2, respectively, and let $w_{(1)i}, w_{(2)i}$ denote the phase one and two sampling weights for an element i. Let $\mathbf{x}_i$ be a vector of J_x auxiliary variables known for all $i \in U$, where $J_x \geq 2$ since $\mathbf{x}_i$ will be assumed to contain an intercept. Let $\mathbf{z}_i$ be a vector of J_z variables observed for all $i \in s_1$. Finally, let y_i denote the variable of interest collected for all $i \in s_2$. The target of the estimation procedure is $\bar{y}_N$, the population mean of the y_k.

The regression estimator considered here is the same as in Fuller (1998), or

$$\widehat{\bar{y}}_{\text{reg}(2)} = \bar{y}_{\pi(2)} + \widehat{\boldsymbol{\beta}}_{y|xz(2)}^T \begin{pmatrix} \bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi(2)} \\ \widehat{\bar{\mathbf{z}}}_{\text{reg}(1)} - \bar{\mathbf{z}}_{\pi(2)} \end{pmatrix}, \tag{1}$$

with $\bar{y}_{\pi(2)} = \sum_{s_2} w_{(2)i} y_i$ the two-phase expansion estimator of $\bar{y}_N$ and analogously for $\bar{\mathbf{x}}_{\pi(2)}, \bar{\mathbf{z}}_{\pi(2)}$, and

$$\widehat{\bar{\mathbf{z}}}_{\text{reg}(1)} = \bar{\mathbf{z}}_{\pi(1)} + \widehat{\boldsymbol{\beta}}_{z|x(1)}^T \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi(1)}\right), \tag{2}$$

the phase one regression estimator of $\bar{\mathbf{z}}_N$ where $\bar{\mathbf{z}}_{\pi(1)} = \sum_{s_1} w_{(1)i} \boldsymbol{z}_i$ and $\bar{\mathbf{x}}_{\pi(1)}$ are phase one expansion estimators. In (1) and (2), $\widehat{\boldsymbol{\beta}}_{y|xz(2)}$ is a vector of design-weighted regression coefficients for the y_i on $\mathbf{x}_i, \mathbf{z}_i$ fitted on s_2, and $\widehat{\boldsymbol{\beta}}_{z|x(1)}$ is a matrix with each column containing the design-weighted regression coefficients for one of the variables in $\mathbf{z}_i$ on $\mathbf{x}_i$ fitted on s_1. Note that the estimator (1) can be expressed as a weighted sum over the phase two sample, $\widehat{\bar{y}}_{\text{reg}(2)} = \sum_{s_2} w^*_{(2)i} y_i$, where $w^*_{(2)i}$ denotes the phase two regression weight for element i.

Under some regularity conditions not further discussed here, we can show that the asymptotic variance of $\widehat{\bar{y}}_{\text{reg}(2)}$ is equal to

$$\begin{aligned} \text{AV}(\widehat{\bar{y}}_{\text{reg}(2)}) &= \text{Var}\left(\bar{e}_{(y|xz)\pi(2)}\right) \\ &\quad + \boldsymbol{\beta}_{y|xz,z}^T \text{Var}\left(\bar{\boldsymbol{e}}_{(z|x)\pi(1)}\right) \boldsymbol{\beta}_{y|xz,z} \\ &\quad + 2\boldsymbol{\beta}_{y|xz,z}^T \text{Cov}\left(\bar{e}_{(y|xz)\pi(2)}, \bar{\boldsymbol{e}}_{(z|x)\pi(1)}\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}_{y|xz,z}$ is the portion of the vector of regression coeffients for y_i on $\boldsymbol{x}_i, \boldsymbol{z}_i$ fitted on the population that corresponds to the $\boldsymbol{z}_i$ covariates, $\bar{e}_{(y|xz)\pi(2)}$ is the phase two expansion estimator for the residuals $e_{(y|xz),i} = y_i - (\boldsymbol{x}_i, \boldsymbol{z}_i)\boldsymbol{\beta}_{y|xz}$, and $\bar{\boldsymbol{e}}_{(z|x)\pi(1)}$ is the vector of phase one expansion estimators for the residuals $\boldsymbol{e}_{(z|x),i}$ obtained from population regressions of each of the elements of $\boldsymbol{z}_i$ on $\boldsymbol{x}_i$. This follows from the fact that

$$\begin{aligned}\widehat{\bar{y}}_{\mathrm{reg}(2)} - \bar{y}_N &\approx \bar{e}_{(y|xz)\pi(2)} - \bar{e}_{(y|xz)N} \\ &\quad + \boldsymbol{\beta}^T_{y|xz,z}(\bar{\boldsymbol{e}}_{(z|x)\pi(1)} - \bar{\boldsymbol{e}}_{(z|x)N})\end{aligned} \tag{4}$$

as can be checked using standard Taylor series arguments.

To construct an estimator for $\mathrm{AV}(\widehat{\bar{y}}_{\mathrm{reg}(2)})$, we start from a generic replication method that produces design consistent variance and covariance estimators for phase one and phase two expansion estimators. We will specify the replication method further below. Suppose for now that we have a method to construct R sets of phase one and phase two replicate samples based on s_1 and s_2. Let $w^{(r)}_{(1)i}, w^{(r)}_{(2)i}$ represent the phase one and two expansion weights for the ith element in replicate r. For some variable a_i, let $\bar{a}^{(r)}_{\pi(1)} = \sum_{s_1} w^{(r)}_{(1)i} a_i, r = 1, \ldots, R$ represent the replicate estimators computed for expansion estimator $\bar{a}_{\pi(1)}$, and similarly, let $\bar{a}^{(r)}_{\pi(2)} = \sum_{s_2} w^{(r)}_{(2)i} a_i$ represent the rth replicate for $\bar{a}_{\pi(2)}$. Define

$$\widehat{V}(\bar{a}_{\pi(1)}) = c_R \sum_{r=1}^{R} \left(\bar{a}^{(r)}_{\pi(1)} - \overline{\bar{a}^{R}_{\pi(1)}}\right)^2,$$

where $\overline{\bar{a}^{R}_{\pi(1)}} = \frac{1}{R}\sum_{r=1}^{R} \bar{a}^{(r)}_{\pi(1)}$, and similarly for $\widehat{V}(\bar{a}_{\pi(2)})$. Finally, define

$$\begin{aligned}&\widehat{C}(\bar{a}_{\pi(1)}, \bar{b}_{\pi(2)}) = \\ &\quad c_R \sum_{r=1}^{R} \left(\bar{a}^{(r)}_{\pi(1)} - \overline{\bar{a}^{R}_{\pi(1)}}\right)\left(\bar{b}^{(r)}_{\pi(2)} - \overline{\bar{b}^{R}_{\pi(2)}}\right)\end{aligned} \tag{5}$$

for two expansion estimators $\bar{a}_{\pi(1)}, \bar{b}_{\pi(2)}$. We assume that $\widehat{V}(\bar{a}_{\pi(1)}), \widehat{V}(\bar{a}_{\pi(1)})$ and $\widehat{C}(\bar{a}_{\pi(1)}, \bar{b}_{\pi(2)})$ are design consistent for $\mathrm{Var}(\bar{a}_{\pi(1)}), \mathrm{Var}(\bar{a}_{\pi(2)})$ and $\mathrm{Cov}(\bar{a}_{\pi(1)}, \bar{b}_{\pi(2)})$, respectively.

The replicates for $\widehat{\bar{y}}_{\mathrm{reg}(2)}$ are constructed as follows:

$$\widehat{\bar{y}}^{(r)}_{\mathrm{reg}(2)} = \bar{y}^{(r)}_{\pi(2)} + \widehat{\boldsymbol{\beta}}^{(r)T}_{y|xz(2)} \begin{pmatrix} \bar{\mathbf{x}}_N - \bar{\mathbf{x}}^{(r)}_{\pi(2)} \\ \widehat{\bar{\mathbf{z}}}^{(r)}_{\mathrm{reg}(1)} - \bar{\mathbf{z}}^{(r)}_{\pi(2)} \end{pmatrix},$$

where

$$\widehat{\bar{\mathbf{z}}}^{(r)}_{\mathrm{reg}(1)} = \bar{\mathbf{z}}^{(r)}_{\pi(1)} + \widehat{\boldsymbol{\beta}}^{(r)T}_{z|x(1)} \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}^{(r)}_{\pi(1)}\right),$$

and the $\widehat{\boldsymbol{\beta}}^{(r)}_{y|xz(2)}, \widehat{\boldsymbol{\beta}}^{(r)}_{z|x(1)}$ are regression coefficients computed for each of the R replicates. We define

$$\widehat{V}(\widehat{\bar{y}}_{\mathrm{reg}(2)}) = c_R \sum_{r=1}^{R} \left(\widehat{\bar{y}}^{(r)}_{\mathrm{reg}(2)} - \overline{\widehat{\bar{y}}^{R}_{\mathrm{reg}(2)}}\right)^2 \tag{6}$$

as the replicate variance estimator for the two-phase regression estimator $\widehat{\bar{y}}_{\mathrm{reg}(2)}$. Note that the replicates for the regression estimator can be written as $\widehat{\bar{y}}^{(r)}_{\mathrm{reg}(2)} = \sum_{s_2} w^{*(r)}_{(2)i} y_i$ for a set of replicate regression weights $w^{*(r)}_{(2)i}$. Hence, it is sufficient to have access to a file containing the phase two data and replicate regression weights to be able to compute variance estimator (6).

Under some regularity conditions not further explored here, it is possible to show that $\widehat{V}(\widehat{\bar{y}}_{\mathrm{reg}(2)})$ is design consistent for $\mathrm{AV}(\widehat{\bar{y}}_{\mathrm{reg}(2)})$ in (3). We briefly motivate this result here. Assuming that the replication regression coefficients $\widehat{\boldsymbol{\beta}}^{(r)}_{y|xz(2)}, \widehat{\boldsymbol{\beta}}^{(r)}_{z|x(1)}$ are consistent for $\boldsymbol{\beta}_{y|xz(2)}, \boldsymbol{\beta}_{y|xz(2)}$, respectively, it follows from (4) that

$$\begin{aligned}\widehat{\bar{y}}^{(r)}_{\mathrm{reg}(2)} - \overline{\widehat{\bar{y}}^{R}_{\mathrm{reg}(2)}} &\approx \bar{e}^{(r)}_{(y|xz)\pi(2)} - \overline{\bar{e}^{R}_{(y|xz)\pi(2)}} \\ &\quad + \boldsymbol{\beta}^T_{y|xz,z}(\bar{\boldsymbol{e}}^{(r)}_{(z|x)\pi(1)} - \overline{\bar{\boldsymbol{e}}^{R}_{(z|x)\pi(2)}})\end{aligned}$$

and hence,

$$\begin{aligned}\widehat{V}(\widehat{\bar{y}}_{\mathrm{reg}(2)}) &\approx \widehat{V}(\bar{e}_{(y|xz)\pi(2)}) \\ &\quad + \boldsymbol{\beta}^T_{y|xz,z} \widehat{V}(\bar{\boldsymbol{e}}_{(z|x)\pi(1)}) \boldsymbol{\beta}_{y|xz,z} \\ &\quad + 2\boldsymbol{\beta}^T_{y|xz,z} \widehat{C}(\bar{e}_{(y|xz)\pi(2)}, \bar{\boldsymbol{e}}_{(z|x)\pi(1)}).\end{aligned} \tag{7}$$

Since each of the replication estimators on the right-hand side are assumed consistent for their respective variance term, we conclude that $V(\widehat{\bar{y}}_{\mathrm{reg}(2)})$ is also consistent for $\mathrm{AV}(\widehat{\bar{y}}_{\mathrm{reg}(2)})$.

As shown in (3), the asymptotic variance of $\widehat{\bar{y}}_{\text{reg}(2)}$ contains a covariance term between the residuals of both regressions, and hence a general replication method for two-phase regression estimators needs to be able to estimate this covariance term. By assuming that our chosen variance replication method is able to estimate covariances of expansion estimators between phases as in (5), the method indeed satisfies this requirement. In many regression estimation situations (see Fuller, 1998), the covariance term in (3) will be equal to 0, so that the replication method does not need to be able to estimate covariances for variables across phases.

4. Simulation Results

We conducted a simulation to study the performance of the proposed variance estimator. For this purpose, finite populations of size $N = 10,000$ were created using the model

$$y_i = 10 + 5x_i - 5z_i + \varepsilon_{y,i},$$

where $\varepsilon_{y,i} \sim N(0, \sigma_y^2)$, and we will investigate two levels for σ_y^2. We also consider three cases for the relationship between z_i and x_i. In the first two cases, z_i and x_i are linearly related through the model

$$z_i = 1 - x_i + \varepsilon_{z,i},$$

with $x_i \sim U(0,1)$ and $\varepsilon_{z,i} \sim N(0, \sigma_z^2)$, for two levels of σ_z^2. In the third case, z_i and x_i are linearly independent and are both generated as $U(0,1)$.

By crossing the cases for the model for y_i with those for the model for z_i, we obtain six different scenarios for the overall population model, which we will identify by the coefficients of determination of both models, R_y^2 and R_z^2. Specifically, we varied the model for z_i and the model variances so that the six cases correspond to the combinations (R_y^2, R_z^2) with $R_y^2 = 0.25$ or 0.75, and $R_z^2 = 0, 0.25$ or 0.75.

Two-phase samples were drawn from each of the populations, with simple random sampling without replacement in both phases. The sample sizes were $n_1 = 2,000$ for phase one, and $n_2 = 20$ or 200 for phase two.

For each sample, the regression estimator (1) was computed, as well as two "delete-a-group" jackknife variance estimators for $R = 8$ and 16. These jackknife estimators were constructed by first deleting a $1/R$ fraction of the sample observations in each phase for each replicate, computing the replicate sample means and regressions based on the remaining $(R-1)/R$ fraction of the sample, and then proceeding as in Section 3 (see Kott, 2001). For comparison purposes, we also computed the linearized variance estimator corresponding to $\text{AV}(\widehat{\bar{y}}_{rmreg(2)})$ in (3) but with all unknown quantities replaced by sample-based estimators.

Table 1 displays the simulated bias of the jackknife variance estimators $\widehat{V}_{JK,8}$ and $\widehat{V}_{JK,16}$ and the linearized variance estimator $\widehat{V}_L$, as a percentage of the true (simulated) variance $\text{Var}(\widehat{t}_{y\text{reg}(2)})$, for the six populations and two phase two sample sizes considered. Both the jackknife and the linearized estimators are severely biased for the smallest sample size, with the former overestimating and the latter underestimating the true variance. At the larger sample size, the same pattern of over and underestimating is still visible, but the magnitude of the bias is reduced to less than six percent in all cases. Increasing the number of replicates improves the bias properties of the jackknife estimator, and at the larger phase two sample size considered, the bias appears to be at an acceptable level for most practical applications.

Table 2 displays the simulated mean squared error (MSE) of the variance estimators. To make comparison easier, these results were scaled by the MSE of the design variance estimators of $\bar{y}_{\pi(2)}$. These results show that the jackknife estimators are more variable than the linearized estimator by an order of magnitude. This is not surprising, since the former are based on much smaller degrees of freedom than the latter. Nevertheless, it is a cause for concern for group-based replication methods and indicates that the number of replicates should not be chosen too small to avoid unduly variable estimators.

	$\frac{\widehat{V}_{JK,16}}{\mathrm{Var}(\bar{y}_{\pi(2)})} - 1$		$\frac{\widehat{V}_{JK,8}}{\mathrm{Var}(\bar{y}_{\pi(2)})} - 1$		$\frac{\widehat{V}_{L}}{\mathrm{Var}(\bar{y}_{\pi(2)})} - 1$	
Population	$n_2 = 200$	$n_2 = 20$	$n_2 = 200$	$n_2 = 20$	$n_2 = 200$	$n_2 = 20$
$R_y^2 = 0.75$ $R_z^2 = 0.75$	2.62	26.76	3.01	29.58	-2.78	-18.32
$R_y^2 = 0.75$ $R_z^2 = 0.25$	5.16	24.95	5.84	25.15	-1.39	-19.30
$R_y^2 = 0.75$ $R_z^2 = 0$	5.75	22.34	5.08	27.32	-2.61	-20.95
$R_y^2 = 0.25$ $R_z^2 = 0.75$	4.03	25.66	4.90	26.86	-1.33	-19.01
$R_y^2 = 0.25$ $R_z^2 = 0.25$	4.54	22.41	5.45	25.23	-1.06	-20.42
$R_y^2 = 0.25$ $R_z^2 = 0$	4.42	23.31	4.52	23.13	-1.71	-20.60

Table 1: Simulated relative bias of "delete-a-group"jackknife variance estimators for $R = 8$ and 16 groups and of linearized variance estimator (in percent).

5. Jackknife Variance Estimation for the NRI

The replication method used in NRI variance estimation is a form of "delete-a-group jackknife" (Kott, 2001). The 1997 NRI is the phase one sample, and we will discuss the 2001 NRI as the phase two sample, but the same procedure is to be used in all subsequent years. For both the 1997 and the 2001 NRI, we suppose that we have completed the estimation process described in Section 2, so that all the necessary pseudo-points have been created and the estimation weights are calibrated for the auxiliary information at the population level and, in the case of the 2001 NRI, for a set of estimates in the 1997 NRI.

Let $w_{(1)i}, w_{(2)i}$ represent the original design weights for a point i in phase one (1997 NRI) or phase two (2001 NRI), respectively, and let $w^*_{(1)i}, w^*_{(2)i}$ represent its fully calibrated weights. Hence, for any variable y_i, estimates for its 1997 and 2001 totals are computed as $\widehat{t}_{y(1)} = \sum_{s_1} w^*_{(1)i} y_i$ and $\widehat{t}_{y(2)} = \sum_{s_2} w^*_{(2)i} y_i$, respectively (NRI weights are designed for totals, not means). The goal of the variance estimation procedure for the 2001 NRI is to construct R new sets of weights $w^{*(r)}_{(2)i}$, from which a user of the NRI data is able to compute replicate estimates $\widehat{t}^{(r)}_{y(2)} = \sum_{s_2} w^{*(r)}_{(2)i} y_i$ for any variable y_i. The user can then calculate a variance estimate for $\widehat{t}_{y(2)}$ by applying the $\widehat{t}^{(r)}_{y(2)}, r = 1, \ldots, R$ in variance formula (6).

Our procedure has a few unique features that distinguish it from other "delete-a-group" jackknife implementations. In most implementations, a weight of zero ("delete") is assigned to the points in a group in each replicate. In our approach, we are giving those points a small but non-zero weight. This is done to preserve the full set of pseudo-points in each replicate, and avoid the risk of obtaining empty calibration cells for rare control categories for some replicate samples. We will also start from the calibrated weights $w^*_{(1)i}, w^*_{(2)i}$ instead of the design weights, in order to minimize the computation burden of the procedure. The major steps in the construction of the replicate weights $w^{*(r)}_{(2)i}$ are summarized here. For full details, see Fuller et al. (2003).

As a first step, R replicate sets of phase one weights $w^{*(r)}_{(1)i}$ are constructed for the 1997 NRI. The sample is divided into non-overlapping groups $G_{1,r}, r = 1, \ldots, R$ by ordering the segments geographically and then using systematic sampling to create groups of approximately

Population	$\frac{\text{MSE}(\widehat{V}_{JK,16})}{\text{MSE}(\widehat{V}(\bar{y}_{\pi(2)}))}$		$\frac{\text{MSE}(\widehat{V}_{JK,8})}{\text{MSE}(\widehat{V}(\bar{y}_{\pi(2)}))}$		$\frac{\text{MSE}(\widehat{V}_{L})}{\text{MSE}(\widehat{V}(\bar{y}_{\pi(2)}))}$	
	$n_2 = 200$	$n_2 = 20$	$n_2 = 200$	$n_2 = 20$	$n_2 = 200$	$n_2 = 20$
$R_y^2 = 0.75$ $R_z^2 = 0.75$	1.41	0.78	2.94	1.14	0.10	0.12
$R_y^2 = 0.75$ $R_z^2 = 0.25$	1.32	0.54	2.82	0.92	0.08	0.10
$R_y^2 = 0.75$ $R_z^2 = 0$	1.68	0.66	3.32	1.05	0.09	0.11
$R_y^2 = 0.25$ $R_z^2 = 0.75$	9.07	4.82	19.01	7.74	0.61	0.76
$R_y^2 = 0.25$ $R_z^2 = 0.25$	9.05	4.29	18.41	7.61	0.58	0.81
$R_y^2 = 0.25$ $R_z^2 = 0$	8.94	5.13	19.09	6.76	0.60	0.83

Table 2: Simulated mean squared error of "delete-a-group"jackknife variance estimators for $R = 8$ and 16 groups and of linearized variance estimator, scaled by the mean squared error of the design variance of the expansion estimator.

equal size. The geographic ordering is used to reflect the stratification of the original sample.

We define the constant $a_R = \sqrt{(R-1)/R}$. Initial weights $w_{(1)i,0}^{*(r)}$ for jackknife replicate r are obtained as follows. For $i \in G_{1,r}$,

$$w_{(1)i,0}^{*(r)} = \begin{cases} w_{(1)i}^{*} - a_R\, w_{(1)i} & \text{if } w_{(1)i}^{*} \geq w_{(1)i} \\ (1 - a_R) w_{(1)i}^{*} & \text{otherwise,} \end{cases}$$

and for $i \notin G_{1,r}$,

$$w_{(1)i,0}^{*(r)} = \begin{cases} w_{(1)i}^{*} + \frac{a_R}{R-1}\, w_{(1)i} & \text{if } w_{(1)i}^{*} \geq w_{(1)i} \\ (1 + \frac{a_R}{R-1}) w_{(1)i}^{*} & \text{otherwise.} \end{cases}$$

However, in order to account for the calibration of phase one estimates to population controls, we set $w_{(1)i,0}^{*(r)} = w_{(1)i}^{*}$ for all the replicates if the point i is in a category with external controls, regardless of their group. In the 1997 NRI, this includes points with a federal ownership classification and those located in Census Water. Note that data are not collected for federal and Census Water points, but they are kept in the dataset so that the surface area of the country is fully represented.

For points that fall in those control categories, the final replicate weight is kept the same as the initial replicate weight, $w_{(1)i}^{*(r)} = w_{(1)i,0}^{*(r)}$, since that weight is already fully calibrated. For the points that do not fall in one of the control categories, the initial replicate weights are calibrated (through a raking procedure) for a set of additional control acreages available for the previous years, resulting in the final replicate weights $w_{(1)i}^{*(r)}$.

Next, initial 2001 NRI replicate weights $w_{(2)i,0}^{*(r)}$ are constructed using a procedure similar to that described for the 1997 NRI. Since the 2001 NRI sample was selected through an unequal-probability stratified sample from the 1997 NRI, the 2001 NRI sample is sorted by its selection strata and by geography within each stratum. The groups $G_{2,r}$, $r = 1, \ldots, R$ are then again selected by systematic sampling. The initial replicate weights are obtained by setting

$$w_{(2)i,0}^{*(r)} = \begin{cases} w_{(2)i}^{*} - a_R\, w_{(2)i} & \text{if } w_{(2)i}^{*} \geq w_{(2)i} \\ (1 - a_R) w_{(2)i}^{*} & \text{otherwise} \end{cases}$$

for $i \in G_{2,r}$, and

$$w_{(2)i,0}^{*(r)} = \begin{cases} w_{(2)i}^{*} + \frac{a_R}{R-1}\, w_{(2)i} & \text{if } w_{(2)i}^{*} \geq w_{(2)i} \\ (1 + \frac{a_R}{R-1}) w_{(2)i}^{*} & \text{otherwise} \end{cases}$$

for $i \notin G_{2,r}$, except if point i is in one of the categories with the same external controls as for

the the 1997 NRI. In the latter case, $w^{*(r)}_{(2)i,0} = w^{*}_{(2)i}$ as before, and we let the final replicate weight $w^{*(r)}_{(2)i} = w^{*}_{(2)i}$ as well.

For the points that do not fall in one of the categories with external controls, the initial 2001 NRI replicate weights $w^{*(r)}_{(2)i,0}$ for each r are raked to a set of state-level estimates obtained using the corresponding final 1997 NRI replicate weights $w^{*(r)}_{(1)i}$. The categories for which the phase two estimates are calibrated to the phase one are the same as those used in the construction of the original weights $w^{*}_{(2)i}$, and include several broad landuse categories as well as a wetland classification. At the end of this raking step, we obtain the final replicate weights $w^{*(r)}_{(2)i}$, which are appended to the NRI 2001 dataset for the purpose of variance calculation.

In this procedure, no attempt is made to "nest" the 2001 replicates inside the 1997 replicates, which significantly simplifies the construction of the replicate samples. By calibrating each phase two replicate to a phase one replicate, the procedure successfully incorporates both phase one and phase two variability into the resulting phase two replicates. However, it fails to capture any covariances across phases. This is reasonable in this case, since by construction of the calibration steps in the NRI, the cross-phase covariance term in the variances (see (3) above) are expected to be close to 0.

6. Conclusion

In this article, we have proposed a replication variance estimation procedure for fully calibrated two-phase regression estimators. The procedure is quite general, and can be used in many multi-phase survey context in which complex estimation procedures are used. Simulation results show that the procedure has low bias but the number of replicate samples should not be taken too low. One of the main advantages of this procedure is that, once replicate weights have been generated, variance estimates can be computed using only the phase two sample data.

The procedure is being implemented for the variance estimation of the annual NRI surveys, and initial investigation of the resulting estimates is promising. We are currently using $R = 30$ replicates, which represents a trade-off between the computational burden and the need to achieve sufficiently stable estimates. On-going research focuses on trying to simplify the calibration and raking computations for the replicate samples.

References

Estevao, V. M. and C.-E. Särndal (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics 18*, 233–255.

Fuller, W. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica 8*, 1153–1164.

Fuller, W. A. (1999). Estimation procedures for the United States National Resources Inventory. In *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, pp. 39–44.

Fuller, W. A. (2003a). The 2001 NRI estimation procedures. Unpublished manuscript.

Fuller, W. A. (2003b). Sample selection for the 2000 NRI–2004 NRI surveys. Unpublished manuscript.

Fuller, W. A., W. J. Kennedy, D. Schnell, G. Sullivan, and H. J. Park (1986). *PC-CARP*. Ames, IA: Statistical Laboratory, Iowa State University.

Fuller, W. A., J. D. Opsomer, and X. Li (2003). Variance estimation for the 2001 NRI. Unpublished manuscript.

Kim, J. K. and R. R. Sitter (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica 13*, 641–653.

Kott, P. S. (2001). The delete-a-group jackknife. *Journal of Official Statistics 17*, 521–526.

Li, X. and J. D. Opsomer (2003). A comparison of two regression estimators for two-

phase sampling. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA. American Statistical Association. To appear.

Nusser, S. M. and J. J. Goebel (1997). The national resources inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics 4*, 181–204.

Rao, J. N. K. and R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika 82*, 453–460.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association 92*, 780–787.