Estimation for Longitudinal Surveys with Repeated Panels of Observations

Jason Legg, Wayne A. Fuller, and Sarah M. Nusser
Center for Survey Statistics and Methodology
Iowa State University

Abstract

We consider a longitudinal study composed of a first-phase sample with multiple subsets of these units selected for observation over time. Such a design is used for the National Resource Inventory, where a core panel of segments is observed yearly and annual supplements are selected using a rotation design. As observations are taken over time, there is a dependency in the data that can be exploited in estimation. We use an estimated generalized least squares (EGLS) approach that utilizes the estimated time dependency to improve estimation of level and change relative to direct survey estimators. Because longitudinal studies often involve a large number of variables and the output of such studies is a dataset with weights for end users, we provide a s consistent jackknife replication variance method for our EGLS estimator. This approach relies on having a consistent jackknife variance estimator for the first-phase sample. The National Resource Inventory will serve as the motivating example for this work.

Introduction

Longitudinal surveys are surveys in which data are collected at more than one point in time. Examples of longitudinal surveys are panel surveys, which are surveys where observations are taken on the same unit more than once over time. Longitudinal panel surveys have gained in use and importance in decision making over the past 30 years because they provide efficient estimators of changes over time (Kasprzyk 1989). Today, longitudinal surveys such as the Forest Inventory and Analysis National Program, Current Population Survey, and National Health Inventory Survey are used by policymakers to evaluate past decisions and develop future policies. The defining characteristic of panel surveys is that repeated observations are made on the same unit over time. The repetition of observations on the same units in the survey exists for studying changes over time in the responses from the observation units.

Design and estimation for longitudinal panel surveys provide two challenges. The design of a panel survey is composed of two components. The first is a probability mechanism for selecting a collection of units to observe at some point over a series of surveys. The second is a probability mechanism for assigning the selected units to groups that will be observed at a specific time. The choice of the longitudinal observation scheme is critical to achieving the objectives of the survey. For estimation, we need to determine how to utilize the time dependency in the data as well as how to build estimators that incorporate the longitudinal design structures.

We consider a class of longitudinal panel surveys in which a large, first-phase sample is selected and several second-phase samples are selected from the first large sample and observed at different points in time. We will focus on the estimation problem for two-phase longitudinal designs and propose a model and estimator for means and totals under these designs.

If analysis objectives of a longitudinal panel survey are known, models and estimation schemes can be constructed to satisfy those goals. However, many large-scale longitudinal surveys produce data for use by end-users where the models and parameters of interest are not completely known to the designers of the survey. When a survey's ultimate use is not completely prespecified, the result of the survey may be a dataset with estimation weights that can be used in the construction of estimators built from a common form, for example from Horvitz-Thompson total estimators. Similarly, replication weights for estimating variances of these estimators are often provided. With the outputted dataset, end users control the estimators and are not restricted to predetermined estimators.

We consider a cell-mean model where the auxiliary information is a set of indicator variables and the time dependency of observations on the same unit are incorporated through the correlation ma-

trix. Given the cell mean structure, we construct the estimated generalized least squares (EGLS) estimator. EGLS is a method for estimating the minimum variance estimator for parameters in a linear model. We also present a replication variance consistency result.

The National Resources Inventory (NRI) is a longitudinal survey that changed from a 5-year panel study into a yearly supplemented panel survey in 2000. Since the time between observations and the number of observations at each time decreased under the new design, there is interest in utilizing the time dependency in the dataset to improve the efficiency of yearly estimates and estimates of change over time.

Two-phase Samples

We consider longitudinal surveys with a first-phase sample containing all of the units that will be studied over time. In two-phase sampling, we begin by first selecting a large sample of units, which we call the first-phase sample and we observe a set of auxiliary variables for units in the first-phase. A second-phase sample is selected from the population and often depends on the first-phase observations. The second-phase sample is often a sub-sample from the first-phase sample, and we restrict our discussion to this type of two-phase sampling. In the second-phase sample, we observe both the auxiliary variables and the response variables. The auxiliary variables are often less expensive to observe than the response variables, and the auxiliary variables are often chosen to be related to the response variables.

To define a two-phase sample where the second-phase sample is a subsample of the first-phase sample, let A_1 be the first-phase sample of size n, drawn from some finite population, denoted by $\mathcal{F}_N$, with size N. The design for the first-phase sample is denoted by $p_1(\cdot)$, with associated inclusion probabilities $\pi_{1i} = Pr[\text{unit } i \in A_1]$ for $i = 1, 2, \ldots, N$. From the first-phase sample, we select the second-phase sample A_2, with design $p_2(\cdot|A_1)$ and associated conditional inclusion probabilities of $\pi_{2i|1i} = Pr[\text{unit } i \in A_2 \text{ given unit } i \in A_1]$ for $i = 1, 2, \ldots, n$.

For the longitudinal designs we will discuss, the second-phase samples will be subsamples from the first-phase sample. To support development of statistical theory, we will view the second-phase design of these surveys as being defined by a partitioning mechanism that divides the first-phase into P groups, or panels, denoted by A_{2_p} with $p = 1, 2, \ldots, P$. A panel, or union of panels, is assigned by the longitudinal design structure to be observed at specific times. For a longitudinal survey, $p_2(\cdot|A_1)$ defines the probability that each unit $i \in A_1$ is assigned to A_{2_p} for $p = 1, 2, \ldots, P$. For example, $p_2(\cdot|A_1)$ may provide a multinomial probability for each unit in A_1. Let n_p be the second-phase sample size for A_{2_p}.

The described class of samples includes many common longitudinal sample designs. The pure panel, in which each sampled unit is observed at every time point, can be thought of as having a second-phase sample where the entire first-phase sample is the second-phase sample. Selecting a pure rotating panel survey, in which a single panel is observed at each time point and then not reobserved until all panels have been observed, can be thought of as first selecting a first-phase sample composed of all of the units we will observe over time, then partitioning the first-phase sample into P panels and observing one panel at each time point until all P panels are observed, then repeating through the panels. A supplemented panel design combines the notions of a pure panel and pure rotating sample by partitioning a first-phase sample into a core set of units to be observed at every time point and a set of rotating panels that are cycled through as in a rotating panel design. Other combinations of panels such as observing two panels at a time point and replacing one at each time point also fall into the class of designs under consideration.

National Resources Inventory

The National Resources Inventory (NRI) has its roots in the Natural Resource Conservation Service's (NRCS) efforts to monitor soil erosion which dates back to the 1930s. Over time, NRCS objectives grew and in 1982, the NRI survey was created to address a broader set of goals and natural resource concerns (Nusser and Goebel 1997). The objectives of the NRI are to monitor conditions and trends of soil, water, and related natural resources on non-federal lands. These objectives were developed in response to the increased importance of addressing agro-environmental and ecological problems (Goebel 1998). More recently the NRI has expanded its wetlands variables and has focused on observing changes on the land with respect to urban sprawl.

The original longitudinal design for the NRI was a pure panel observed every five years. Roughly

300,000 segments were observed every five years. The 300,000 segments observed in 1997 form the first-phase sample for our discussion since it was partitioned into the panels for the continuous inventory sample.

The 1997 Foundation NRI sample is a stratified two-stage area sample. In a typical Public Land Survey System state, the strata are two mile by six mile areas (one-third of a township). The area segments (sometimes referred to as primary sampling units or PSUs) are typically half-mile by half-mile land areas. Sampling rates vary, but two segments are usually selected within each stratum. Within the selected segments, a secondary sampling unit (SSU) is a sample point. Three sample points are typically selected within each sampled segment. Points were selected using a restricted sampling scheme that guarantees spatial dispersion of selected points in the sample (Nusser and Goebel 1997).

The NRI began an annual inventory in 2000. The decision to move to an annual, or continuous, inventory was made to satisfy the demand by users to have current estimates and to establish a more stable operational base for collecting data. The workload is distributed over each of five year individually rather than making all of the observations at the same time every five years. It is believed that measurement error induced by the data collectors may be reduced under the annual data collection approach (Breidt and Fuller 1999).

The longitudinal design for the continuous inventory is a supplemented panel design. The core, the pure panel component, and supplemental panels, the rotating panel components, are subsamples selected from the 1997 Foundation NRI sample. The second-phase design describes the probabilities used to select the core and supplemented samples from the first-phase 1997 Foundation NRI sample. The core sample contains approximately 41,600 segments (Fuller and Wang 2002). Only the core sample was observed in 2000, with supplements beginning in 2001. The 2001 supplement contains approximately 32,000 segments and each successive supplement is intended to be similar in size. Table 1 shows the NRI data collection for 2000-2003.

A supplemented panel design is a compromise design between the competing objectives of estimating current level and change. The core panel provides the continuous reobservation of units that is often beneficial to change estimation. The supplements are a set of disjoint panels that are rotated in and out of the study. Often only one supplemental panel is observed at each time point and the supplement is reobserved after all of the supplements have been observed. The NRI supplemented panel design's rotating of panels is more complex and occurs more on a land category basis. Specific details of the second-phase sample can be found in Fuller and Wang (2001).

Table 1: 2000-2003 NRI Data Collection

Year:	00	01	02	03
Core	X	X	X	X
Supplement 1		X		
Supplement 2			X	
Supplement 3				X

Generalized Least Squares

For two-phase samples, the basic building block for estimators is the $\pi^\star$-expanded estimator where the $\pi^\star$-expanded estimator for a population total is

$$\hat{t}_{y,\pi^\star} = \sum_{i \in A_2} \pi_i^{\star -1} y_i. \tag{1}$$

The $\pi^\star$-expanded estimator, like the Horvitz-Thompson estimator, is design unbiased and has a design unbiased variance estimator.

Stukel and Kott (1996) examined two two-phase sample estimators in the case where the second phase sample is stratified. Stukel and Kott term the $\pi^\star$-expanded estimator in the stratified case to be the double expansion estimator (DEE). The second estimator is the ratio of two $\pi^\star$-expanded estimators, one for estimating the total for y, the other for estimating the size of the strata. This ratio estimator is termed the reweighted expansion estimator (REE).

An extension to the REE and DEE is the two-phase regression estimator, which makes use of auxiliary variables observed at the first phase (Fuller 2005). Let $\boldsymbol{x}_i$ be a $k \times 1$ auxiliary information vector observed for units in A_1. Let y_i be the response variable observed in A_2 and let the regression estimator for a population mean be

$$\bar{y}_{reg} = \bar{y}_{\pi,A_2} + (\bar{\boldsymbol{x}}_{\pi,A_1} - \bar{\boldsymbol{x}}_{\pi,A_2})\hat{\boldsymbol{\beta}}_2, \tag{2}$$

where

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i \in A_2} \pi_i^{\star -1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\pi,A_2})'(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\pi,A_2}) \right)^{-1} \times \sum_{i \in A_2} \pi_i^{\star -1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\pi,A_2})'(y_i - \bar{y}_{\pi,A_2}), \tag{3}$$

$$(\bar{\boldsymbol{x}}_{\pi,A_2}, \bar{y}_{\pi,A_2}) = \left(\sum_{i\in A_2} \pi_i^{\star -1}\right)^{-1} \sum_{i\in A_2} \pi_i^{\star -1}(\boldsymbol{x}_i, y_i), \tag{4}$$

and

$$\bar{\boldsymbol{x}}_{\pi,A_1} = \left(\sum_{i\in A_1} \pi_{1i}^{-1}\right)^{-1} \sum_{i\in A_1} \pi_{1i}^{-1}\boldsymbol{x}_i. \tag{5}$$

The regression estimator provides a way of using the correlation between observations taken at two different times in the sample. For a longitudinal survey, we have multiple times and panels of observations to combine. Generalized least squares with a cell-mean model combines the different components of the sample.

Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of observations, $\boldsymbol{X}$ is the $n \times k$ matrix of auxiliary variables, $\boldsymbol{\beta}$ is the $q \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors with

$$E(\boldsymbol{\epsilon}) = \boldsymbol{0} \tag{7}$$

and

$$V(\boldsymbol{\epsilon}) = \boldsymbol{V}. \tag{8}$$

Generalized Least Squares (GLS) is an estimation method that provides an estimator for $\boldsymbol{\beta}$. For positive definite $\boldsymbol{V}$ and invertible $\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$, the GLS estimator is

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{9}$$

In practice, the covariance matrix of $\boldsymbol{\epsilon}$ is not known, so a consistent estimator of $\boldsymbol{V}$, $\hat{\boldsymbol{V}}$, is created. Substituting $\hat{\boldsymbol{V}}$ into (9), provides the EGLS estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{EGLS} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{10}$$

For estimation in repeated surveys, we will take $\boldsymbol{Y}$ to be the vector of $\bar{y}_{\pi,pt}$'s, estimators of the population mean for panel p at time t. We will assume $\bar{y}_{\pi,pt}$ is consistent for the population mean at time t. For example, $\bar{y}_{\pi,pt}$ may be the $\pi^{\star}$-expanded estimator ratio estimator. For the NRI, $\bar{y}_{\pi,pt}$ is a ratio estimator that uses the NRI estimation weights. Let y_{it} be the sample mean of y collected on points in segment i with data collected at time t for a variable. Let w_i be the weight associated with segment i. The core and supplements are considered panels and are indexed by p. For the NRI,

$$\bar{y}_{\pi,pt} = \frac{\sum_{i\in A_{2_p}} w_i y_{it}}{\sum_{i\in A_{2_p}} w_i}. \tag{11}$$

The unknown parameter vector $\boldsymbol{\beta}$ will be the $T \times 1$ vector of μ_t, the population means at time point t for $t = 1, 2, \ldots, T$. Since we consider a cell-mean model, $\boldsymbol{X}$ will be the matrix of 0's and 1's linking each $\bar{y}_{\pi,pt}$ to the corresponding μ_t. For $\bar{y}_{\pi,pt}$, the corresponding row in $\boldsymbol{X}$ will have a 1 in the cell that is multiplied by μ_t, and 0 in the remaining cells. We will assume observations from different units are independent and have a constant variance. The constant variance assumption implies that $\boldsymbol{V}$ will be proportional to the correlation matrix of the $\bar{y}_{\pi,pt}$'s. The independence assumption implies that only correlations between $\bar{y}_{\pi,pt}$'s on the same panel will have a nonzero correlation. We will estimate $\boldsymbol{V}$ with a design consistent estimator $\hat{\boldsymbol{V}}$. The described model and assumptions are outlined in Fuller (1990) for general repeated surveys and Fuller and Breidt (1999) for supplemented panel surveys.

For the data in Table 1, the model is

$$\boldsymbol{y}_1 = \begin{pmatrix} \bar{y}_{\pi,00} \\ \bar{y}_{\pi,01} \\ \bar{y}_{\pi,02} \\ \bar{y}_{\pi,03} \\ \bar{y}_{\pi,11} \\ \bar{y}_{\pi,22} \\ \bar{y}_{\pi,33} \end{pmatrix} \quad \boldsymbol{X}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{12}$$

$$\boldsymbol{\mu}_{T_1} = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix},$$

where $\bar{y}_{\pi,pt}$ is the mean estimated as in (11) for our variable of interest in panel p in year t, and μ_t is the population mean of our variable in year t. A panel code of 0 represents the core panel, and the other values of t represent supplements first seen in the year $200t$. We assume that each panel mean has the same constant variance term so the variances of the panel means differ only by the sample sizes. Since the core and supplements were selected to have similar compositions of land covers and uses, the assumption of the same variance for panel means is reasonable. The model assumption that the supplement and core means in a year t estimate the same mean, μ_t is also reasonable since the core and supplements are random samples from the first-phase sample. We also assume the same sample size in each supplement for

convenience. Let

$$r = \frac{n_0}{n_p}, \tag{13}$$

where n_0 is the sample size in the core, and n_p is the sample size in a supplement. We set r=1.294, the approximately ratio of the sample size in the core to the sample size in the supplements. We assume a stationary error process. Let $\rho(l)$ be the correlation between observations on the same unit at a lag of l. The $\bar{y}_{\pi,pt}$'s with the same p have a nonzero covariance term. Panel means from different panels are assumed to be uncorrelated. Under these assumptions,

$$Var(\boldsymbol{y}_1) \propto \begin{pmatrix} 1 & \rho(1) & \rho(2) & \rho(3) & 0 & 0 & 0 \\ \rho(1) & 1 & \rho(1) & \rho(2) & 0 & 0 & 0 \\ \rho(2) & \rho(1) & 1 & \rho(1) & 0 & 0 & 0 \\ \rho(3) & \rho(2) & \rho(1) & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & r & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & r & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & r \end{pmatrix}. \tag{14}$$

To estimate the μ's in the model, we need to first estimate the $\rho(l)$'s in V. Empirical correlations may be used in the V matrix, however we chose to use time series models. The $\rho(l)$'s were estimated using nonlinear least squares for fitting the empirical correlations to theoretical autocorrelation functions of first-order autoregressive processes.

The variance of the EGLS estimator for a year mean is compared to the variance of the sample mean for a year. For changes, the variance of a one-year mean difference from EGLS is compared to the variance of a one-year change in sample core means. We consider two time frames in our comparison: 2000-2003 and 1997-2003. The 1997-2003 NRI data includes replication on the panels. For the 1997-2003 model, the data that has not been reobserved since 1997 is excluded. Results for three NRI variables are presented in Tables 2 and 3. Soil loss has low correlation over time, urban has high correlation over time, and cultivated crop has correlations in between urban and soil loss. The values are in ratios of variances with the EGLS variance in the numerator.

For estimation of year means, the EGLS estimator outperforms the sample year mean. For estimating one-year change, the EGLS estimator is superior to the core estimator when 1997 data are included. The variance reduction is greater for estimating year means than for one-year change. For estimating year means, there is little precision gain from including the first-phase (1997) sample data in the model. The addition of the 1997 data reduces the variance for year means estimators the most for the variables with the low correlations between years. For variables such as range and cropland, the core estimator of change is performing almost as well as the EGLS estimator when 1997 is not included. The inclusion of the 1997 correlations for the panels improves estimation of change. When the supplements begin to repeat, it is likely that the improvements from the inclusion of the first-phase data will decrease. The results indicate that repeated observations on all of the panels, not just the core, reduce variances of the EGLS estimator for change by including additional correlation information.

Table 2: Variance Ratios for Year Estimators

		Estimation Years	
Variable	Data Used	2000	2003
Soil Loss	2000-2003	0.513	0.688
	1997-2003	0.465	0.681
Urban	2000-2003	0.306	0.538
	1997-2003	0.304	0.537
Cultivated	2000-2003	0.364	0.580
Cropland	1997-2003	0.343	0.577

Table 3: Variance Ratios for One-Year Change Estimators

		Change Years	
Variable	Data Used	1997-2000	2003-2002
Soil Loss	2000-2003		0.902
	1997-2003	0.591	0.860
Urban	2000-2003		0.998
	1997-2003	0.537	0.908
Cultivated	2000-2003		0.971
Cropland	1997-2003	0.563	0.895

Replication Variance

We now present a replication variance consistency theorem for a single $\hat{\mu}_d$ estimator using EGLS with our cell-mean model. The consistency may be extended to the entire vector $\hat{\boldsymbol{\mu}}$ by applying the result to the difference of two $\hat{\mu}_d$'s. Let the number of occasions for observations and panels be fixed at T and P respectively. Let the finite population be of size N indexed by i and t, where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ is time. Let the parameters of interest be the set of population means $\bar{Y}_{t,N} = \sum_{i=1}^{N} N^{-1} Y_{it}$. Define the vector of population means as $\bar{y}_N$. Let $E\{\hat{\theta}|\mathcal{F}\}$ and $V\{\hat{\theta}|\mathcal{F}\}$ denote the design expectation and variance of $\hat{\theta}$, where $\mathcal{F}$ is a finite population. Let

the first-phase sample, that is the sample at t=1, be labeled A_1. Every panel is included in A_1.

Let $\mathcal{F}_N$ be a sequence of increasing populations as defined in Fuller (1975). Define

$$\bar{y}_{\pi,p,t,N} = (\sum_{i\in A_1} \pi_{1i}^{-1}\pi_{2ip|1i}^{-1}a_{ip}b_{pt})^{-1} \times \sum_{i\in A_1} \pi_{1i}^{-1}\pi_{2ip|1i}^{-1}a_{ip}b_{pt}y_{it}, \quad (15)$$

where a_{ip} is an indicator for the i^{th} element in the p^{th} panel, b_{pt} is an indicator for the p^{th} panel at the t^{th} time, π_{1i} is the first-phase selection probability, $\pi_{2ip|1i}$ is the conditional probability of selecting the i^{th} element for the p^{th} panel given that the i^{th} element is in A_1, and the subscript of N is suppressed on the right side of (15). Mathematically,

$$a_{ip} = \begin{cases} 1 & \text{if } i^{th} \text{ unit in the } p^{th} \text{ panel} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

and

$$b_{pt} = \begin{cases} 1 & \text{if } p^{th} \text{ panel observed at time } t \\ 0 & \text{otherwise} \end{cases}. \quad (17)$$

We consider disjoint panels, so a_{ip} may be 1 for only one p. The a_{ip} are random indicator variables with a probability distribution defined by the second-phase design. For the proof of the theorem, we assume that the a_i vectors are independent multinomials of size 1. For each i, we have an independent multinomial distribution with probabilities κ_{2ip} for $p = 1, 2, \ldots, P$. The b_{pt} are fixed indicator variables determined by the longitudinal data collection structure. For example, for the core panel in a supplemented panel design, $b_{core,t}$ is 1 for all t.

Let $\bar{y}_{\pi,N}$ be the vector of $\bar{y}_{\pi,p,t,N}$'s. Let

$$\hat{\mu}_N = (X'\hat{V}_N^{-1}X)^{-1}X'\hat{V}_N^{-1}\bar{y}_{\pi,N} \quad (18)$$

be the EGLS solution , where $\boldsymbol{X}$ is the cell mean model matrix and $\hat{V}_N$ is a consistent estimator of the covariance matrix of $\bar{y}_{\pi,N}$. Define

$$\bar{e}_{\pi,p,t,N} = (\sum_{i\in A_1} \pi_{1i}^{-1}\pi_{2ip|1i}^{-1}a_{ip}b_{pt})^{-1} \times \sum_{i\in A_1} \pi_{1i}^{-1}\pi_{2ip|1i}^{-1}a_{ip}b_{pt}e_{it}, \quad (19)$$

where

$$e_{it} = y_{it} - \bar{y}_{t,N}. \quad (20)$$

Note that $\hat{\mu}_N - \bar{y}_N$ is a vector of linear combinations of $\bar{e}_{\pi,p,t,N}$'s.

Further, assume that

$$C_{\pi S} < Nn_{1N}^{-1}\pi_{1i,N} < C_{\pi B} \quad (21)$$

for all N, where n_{1N} is the sample size for the first-phase sample drawn from $\mathcal{F}_N$, and $C_{\pi S}$ and $C_{\pi B}$ are fixed positive constants. Assume that $\pi_{2ip|1i} = \kappa_{2ip} = Pr[a_{ip} = 1]$ are fixed probabilities, $\sum_{p=1}^{P}\kappa_{2ip} = 1$, and $Pr(i \in p_1 \text{ and } i \in p_2) = 0$ for $p_1 \neq p_2$. Note that κ_{2ip} does not depend on a particular first-phase sample A_1. Assume the finite population $\mathcal{F}_N$ is a sample from an infinite population with $4+\delta$, $\delta > 0$, moments.

The next set of assumptions pertain to first-phase estimators. Assume that

$$V\{\hat{T}_{1y}|\mathcal{F}_N\} \leq K_M V\{\hat{T}_{y,SRS}|\mathcal{F}_N\}, \quad (22)$$

where y is any variable with fourth moments, $\hat{T}_{y,SRS}$ is the total estimator for simple random sampling, $\hat{T}_{1y}$ is the Horvitz-Thompson total estimator from the first-phase, and K_M is a fixed constant. Assume that the variance of a first-phase linear estimator of the mean is a symmetric quadratic function, and that

$$n_N V\{N^{-1}\sum_{i\in A_{1N}} \pi_{1i,N}^{-1}y_{iN}|\mathcal{F}_N\} = \sum_{j=1}^{N}\sum_{i=1}^{N}\omega_{ij,N}y_{iN}y_{jN}, \quad (23)$$

where $\omega_{ij,N}$'s satisfy

$$\sum_{i=1}^{N}|\omega_{ij,N}| = O(N^{-1}). \quad (24)$$

The assumption that underlies the replication variance procedure is that of a design consistent replication variance estimator for the first-phase sample. Let this replication variance estimator for a mean be

$$\hat{V}_1\{\bar{y}_{1,HT}\} = \sum_{k=1}^{L} c_k(\bar{y}_{1,HT}^{(k)} - \bar{y}_{1,HT})^2, \quad (25)$$

where $\bar{y}_{1,HT} = \sum_{i\in A_1} N^{-1}\pi_{1i}^{-1}y_i$ is the Horvitz-Thompson mean of y, $\bar{y}_{1,HT}^{(k)}$ is the k^{th} replicate of the estimated mean, L is the total number of replicates, and c_k, $k = 1, 2, \ldots, L$, are constants determined by the replication method and design.

To construct a replicate for $\hat{\mu}$, apply the replication proceduce for the first-phase to units across time

that share the same element identification. That is, if the i^{th} observation is removed to form a replicate, then data for all time points on the i^{th} element is removed. We then replicate $\hat{\mu}$ using the remaining data. The replication variance estimator for $\hat{\mu}$ is

$$\hat{V}_2(\hat{\mu}_d) = \sum_{k=1}^{L} c_k(\hat{\mu}_d^{(k)} - \hat{\mu}_d)^2, \tag{26}$$

for the d^{th} year where

$$\hat{\mu}^{(k)} = (X'\hat{V}_N^{-1}X)^{-1}X'\hat{V}_N^{-1}\bar{y}_{pt,N}^{(k)} \tag{27}$$

are the replicates. The following result establishes the consistency of $\hat{V}_2\{\hat{\mu}_d\}$.

Assume (21), (22), (23), and (24), as well as

$$E\{[(V[\hat{\theta}|\mathcal{F}_N])^{-1}\hat{V}_1\{\hat{\theta}\} - 1]^2|\mathcal{F}_N\} = o(1) \tag{28}$$

for any variable with bounded fourth moments. Also assume that the replicates for the first-phase sample estimator of a total, $\hat{T}_1$, satisfy

$$E\{[c_{kN}(\hat{T}_1^{(k)} - \hat{T}_1)^2]^2|\mathcal{F}_N\} < K_\gamma L_N^{-2}[V\{\hat{T}_1|\mathcal{F}_N\}]^2 \tag{29}$$

uniformly in N for any variable with fourth moments, where K_γ is a fixed constant. Consider replication variance estimation for μ_d defined in (26). Then the replication variance satisfies

$$\begin{aligned}\hat{V}_2(\hat{\mu}_d) &= V\{\hat{\mu}_d|\mathcal{F}_N\} \\ &\quad -N^{-2}\sum_{i=1}^{N}\left\{\sum_p^P \kappa_{2ip}^{-1}(1-\kappa_{2ip})\eta_{dip}^2 \right. \\ &\quad \left. +\sum_{p_1=1}^{P}\sum_{\substack{p_2=1 \\ p_1\neq p_2}}^{P}(-\eta_{dip_1}\eta_{dip_2})\right\} \\ &\quad +o_p(n^{-1}),\end{aligned} \tag{30}$$

where

$$\eta_{dip} = \sum_{t=1}^{T}\lambda_{dpt}b_{pt}e_{it} \tag{31}$$

and λ_{dpt} are the coefficients of $(X'V^{-1}X)^{-1}X'V^{-1}$ corresponding to estimation of the d^{th} year component of $\hat{\mu}$.

Outline of the Proof: Since the κ_{2ip} do not depend on a particular A_1, the sampling procedure can be thought of by first generating an a_i for each element in the population and then drawing a first-phase sample from the population including the second-phase panel identification. By first determining the panel identification, we can condition on the second-phase sample a_{ip}'s in the proof. Conditioning on the second-phase indicators lets us write the estimator in terms of a first-phase estimator as in (15). Since the first-phase estimator has a consistent replication variance estimator, we can apply that consistency to the estimator conditional on the a_{ip}'s. We then obtain the unconditional properties by applying the probabilities from the independent multinomial distributions of the a_i's.

Acknowledgements

This research was supported in part by the USDA Natural Resources Conservation Service cooperative agreement NRCS-683A754122. I would like to thank the consulting contribution of Dr. Jean Opsomer as well as Iowa State CSSM graduate students for computing and psychological support.

References

[1] F. Jay Breidt and Wayne A. Fuller. Design of supplemented panel surveys with application to the national resources inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(4):391–403, December 1999.

[2] Wayne A. Fuller. Analysis of repeated surveys. *Survey Methodology*, 16(2):167–180, December 1990.

[3] Wayne A. Fuller. *Analysis of Survey Data*, chapter Estimation for Multiple Phase Samples. John Wiley and Sons, Inc., 2003.

[4] Wayne A. Fuller and F. Jay Breidt. Estimation for supplemented panels. *Sankhyā: The Indian Journal of Statistics Series B*, 61(1):58–70, 1999.

[5] Wayne A. Fuller, Kevin W. Dodd, Junyuan Wang, and Charles Peterson. Estimation for the 1997 national resources inventory. Technical report, Iowa State University, 2001.

[6] Wayne A. Fuller and Junyuan Wang. Samples for the continuous inventory. Technical report, Iowa State University, 2001.

[7] J. Jeffery Goebel. The national resources inventory and its role in u.s. agriculture. In *Agricultural Statistics 2000: Proceedings of the conference on agricultural statistics organized by the National Agricultural Statistics Service of the US Department of Agriculture, under the auspices of the International Statistical Institute*, 1998.

[8] Daniel Kasprzyk, Greg J. Duncan, Graham Kalton, and M. P. Singh. *Panel Surveys*. John Wiley and Sons, Inc., 1989.

[9] Jae Kwang Kim, Alfredo Navarro, and Wayne A. Fuller. Replication variance estimation for two-phase stratified sampling. *Pending Publication*, 2005.

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

[11] Diane M. Stukel and Phillip S. Kott. Jackknife variance estimation under two-phase sampling: An empirical investigation. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1996.