Sankhyā: The Indian Journal of Statistics
Special Issue on Sample Surveys
1999, Volume 61, Series B, Pt. 1, pp. 58-70

ESTIMATION FOR SUPPLEMENTED PANELS*

By WAYNE A. FULLER
and
F. JAY BREIDT
Iowa State University, Ames

SUMMARY. In a longitudinal survey conducted for k periods some units may be observed for less than k of the periods. Examples include surveys designed with partially overlapping subsamples, a pure panel survey with nonresponse, and a panel survey supplemented with additional samples for some of the time periods. Estimators of the regression type are exhibited for such surveys. An application to special studies associated with the National Resources Inventory is discussed.

1. **Introduction**

Study of the dynamics of populations requires observations made at multiple time points on units in the population. Surveys designed specifically for such study include the Survey of Income and Program Participation (SIPP) conducted by the U.S. Census Bureau and the National Resources Inventory (NRI) conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. The basic NRI is nearly a pure panel survey of certain land areas with a five year observation interval. However, studies on subsamples have been conducted during some intervening years. We define a pure panel survey to be a survey in which the same units are observed at each time point of a survey conducted at more than one time point. A longitudinal survey is a survey conducted at more than one time point with some units observed at more than one time point. The term longitudinal survey is generally used for surveys conducted at more than two points in time with multiple observations on some units planned as part of the survey design.

AMS (1990) *subject classification.* Primary 62D05, secondary 62D10.
Key words and phrases. Longitudinal surveys, environmental monitoring, multiple phase samples, design of surveys.
* This research was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Service and Iowa State University. Computing for the research reported in this paper was done with equipment purchased with funds provided by an NSF SCREMS grant award DMS 9707740.

A rotation survey is one in which a unit is observed for a partial set of time points and is not observed for the remaining set of time points in the study. There are many ways in which the observation pattern can be specified. The Canadian Labour Force Survey and the U.S. Current Population Survey are examples of surveys designed to run continuously in which units rotate into the sample for a fixed period (or periods) and then permanently rotate out of the observation set.

There exists an array of designs combining individuals observed at some time points and individuals observed at all time points of the study set of time points. The book edited by Kasprzyk *et al.* (1989) contains an excellent discussion of various aspects of panel surveys. Duncan and Kalton (1987) and Schreuder, Gregoire, and Wood (1993) discuss different types of repeated surveys and the objectives of such surveys. An estimation procedure for multiple characteristics in rotating surveys has been suggested by Fuller (1990) and implemented by Lent, Miller, and Cantwell (1996). Variance estimation procedures have been proposed for multiple phase samples by Rao and Sitter (1995), Sitter (1997), and Fuller (1998).

We call a survey in which a panel of individuals is observed at every time point and additional individuals are observed at some of the time points, supplemented panels. Such designs are also called split panels. (See Kish (1986)). The simplest such design is a two-phase sample in which the observations at the second of two time points is a subsample of those observed at time one. We shall pay particular attention to designs in which a supplementary sample of the same size but of different individuals is observed at each time point of the study.

The largest fraction of the survey literature on repeated surveys has been devoted to rotating surveys. An early study describing the use of least squares to incorporate information from a previous occasion into the estimate of the current occasion is that of Jessen (1942), (also see Cochran (1942)). Patterson (1950) investigated estimation for rotating samples. Patterson's work was followed by a number of authors, including Eckler (1955), Rao and Graham (1964), Gurney and Daly (1965), Raj (1965), Smith (1978), Wolter (1979), Jones (1980), Huang and Ernst (1981), Breau and Ernst (1983), Kumar and Lee (1983), Singh (1996), and Yansaneh and Fuller (1998).

We review generalized least squares estimation for partial overlap surveys. We then compare some designs for studies directed toward longitudinal dynamics. Because such studies are multiple purpose, we cannot hope to develop a design that is optimal for all objectives. A part of our investigation will be to identify the trade-offs.

2. Supplemented Panel Designs

In actual applications the sample designs and estimation procedures are complex. To simplify the presentation, we shall assume simple random sampling in our discussion. Also we assume simple fractions and exact replicates in our comparison of designs. These simplifications are not a requirement for estimation purposes and designs under consideration for the National Resources Inventory use different rotation patterns for different subpopulations.

To introduce generalized least squares estimation, consider a simple three-period survey in which one-fourth of the units are observed in all three periods and each of the remaining three sets of one-fourth of the units is observed in exactly one of the three periods. Thus, if the total sample size is n, then 0.5 n of the units are observed at each time point. Let (Y_1, Y_2, Y_3) denote the value of a characteristic observed at times one, two, and three respectively. Assume that the correlation between observations at time i and time j on the same element is $\rho_{|i-j|}$. Assume simple random sampling for the selection of all samples.

Let $(\overline{y}_{11}, \overline{y}_{21}, \overline{y}_{31})'$ denote the estimated mean at time one, two and three, of the sample elements that are observed all three periods. Let $(\overline{y}_{12}, \overline{y}_{23}, \overline{y}_{34})'$ denote the sample means for the three periods for the sample elements that are observed once. We call these six estimators the elementary estimators. (See Gurney and Daly (1965)). Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ denote the population means for the three periods. Then we can write

$$\overline{\mathbf{y}} = \mathbf{X}\boldsymbol{\mu} + \mathbf{e} \qquad \ldots(1)$$

where $\overline{\mathbf{y}}' = (\overline{y}_{11}, \overline{y}_{21}, \overline{y}_{31}, \overline{y}_{12}, \overline{y}_{23}, \overline{y}_{34}), \mathbf{X}' = (\mathbf{I}_3, \mathbf{I}_3)$, the covariance matrix of $\mathbf{e}$ is

$$\boldsymbol{\Sigma}_{ee} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 & 0 & 0 \\ \rho_2 & \rho_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} 4n^{-1}\sigma^2 \qquad \ldots(2)$$

and $4n^{-1}\sigma^2$ is the variance of a mean of $n/4$ observations. It follows that the best linear unbiased estimator of $\boldsymbol{\mu}$ using this amount of information is

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3)' = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\overline{\mathbf{y}} \qquad \ldots(3)$$

and

$$\mathbf{V}\{\hat{\boldsymbol{\mu}}\} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}. \qquad \ldots(4)$$

We compare the covariance matrix (4) with the covariance matrix of a pure panel survey in which the same $n/2$ units are observed on all three periods. The best linear unbiased estimator of $\boldsymbol{\mu}$ under the pure panel design is

$$\hat{\boldsymbol{\mu}}_{panel} = (\overline{y}_1, \overline{y}_2, \overline{y}_3)'$$

and the covariance matrix for the pure panel design is

$$\mathbf{V}\{\hat{\boldsymbol{\mu}}_{panel}\} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} 2n^{-1}\sigma^2.$$

In this section we compare designs under the assumption that $\rho_j = \rho^j$; that is, the correlations are those of a first order autoregressive process with parameter ρ. An extended model is considered in Section 3. Table 1 contains variances for several functions of means for ρ ranging from -0.70 to 0.99, relative to the corresponding variance for the pure panel design. If $\rho = 0$, the variance of the best linear unbiased estimator of the three means is the variance of the simple mean at each period. If the total number of elements is n, the variance at each time period is $2n^{-1}\sigma^2$. In the remaining discussion we assume $\sigma^2 = 1$. We compare designs with the same number of observations. In practice it is possible that observations made on previously observed individuals have a different cost than first time observations.

Table 1. Variances of estimators of parametric functions for three-period design with 50% new observations at each time relative to variances of pure panel design

ρ	μ_1	μ_2	$\mu_1 - \mu_2$	$\mu_1 - \mu_3$
−0.70	0.838	0.755	0.674	1.253
−0.50	0.929	0.875	0.768	1.143
0	1.000	1.000	1.000	1.000
0.50	0.929	0.875	1.304	1.143
0.70	0.838	0.755	1.488	1.253
0.90	0.660	0.595	1.773	1.681
0.99	0.520	0.510	1.973	1.951

The variances of the estimated means for the supplemented panel design for the first and last period are the same by the symmetry of the covariances. Thus the variance of $\hat{\mu}_3$ is the same as the variance of $\hat{\mu}_1$. The variance of the middle period is somewhat smaller because the middle observation has one period correlation with the first and third observations on the same element. Also the variance of $\hat{\mu}_2 - \hat{\mu}_3$ is the same as the variance of $\hat{\mu}_1 - \hat{\mu}_2$, by the symmetry of the covariances.

If observations on the same element are correlated ($\rho \neq 0$), the variance of the estimated means for the supplemented panel design is less than $2n^{-1}$. The relative efficiency for means are the same for negative ρ as for positive ρ. The limit correlation of one can occur for characteristics such as age of individuals measured at equal spaced intervals. The lower bound for the variance of the supplemented panel design at $\rho = 1$ is n^{-1} because this is the number of different individuals in the study. Correspondingly, the limit of the relative efficiency for period means of the supplemented panel, relative to the pure panel, is 2.0.

If the correlation is positive, the variance of the difference of two means is smaller for the pure panel survey than for the supplemented panel design. The pure panel design has an efficiency for change approaching twice that of the supplemented panel design as ρ approaches one. This efficiency holds for any multiple period study. However, as seen in Table 2, the variance of a period-to-period change for both designs approaches zero as ρ approaches one. Thus, for example, the variance of a one-period change for the pure panel is only two percent of the variance of the mean if $\rho = 0.99$. Hence, the limit efficiency for change is a ratio of two quantities both approaching zero.

Table 2. Normalized variances of functions of means for pure panel study

ρ	$nV\{\bar{y}_1\}$	$nV\{\bar{y}_2\}$	$nV\{\bar{y}_1 - \bar{y}_2\}$	$nV\{\bar{y}_1 - \bar{y}_3\}$
−0.70	2.00	2.00	6.80	2.04
−0.50	2.00	2.00	6.00	3.00
0	2.00	2.00	4.00	4.00
.50	2.00	2.00	2.00	3.00
.70	2.00	2.00	1.20	2.04
.90	2.00	2.00	0.40	0.76
.99	2.00	2.00	0.04	0.08

The discussion of the variances of estimated changes makes clear the fact that when one has a vector of estimators it is not possible to choose a design that is optimal for every estimator. Two summary criteria that are used in the design of experiments are the trace of the covariance matrix and the determinant of the covariance matrix. The trace is the sum of the characteristic roots of the covariance matrix and the determinant, called the generalized variance, is the product of the roots. In the following discussion, we normalize the covariance matrix by multiplying by n.

Table 3. Roots of normalized covariance matrix for three-period study

Type	ρ	r_1	r_2	r_3	Sum
Pure Panel	−0.7	4.53	1.02	0.45	6.00
Supplemented		2.77	1.35	0.74	4.86
Pure Panel	−0.5	3.69	1.50	0.81	6.00
Supplemented		2.59	1.71	1.16	5.46
Pure Panel	0.0	2.00	2.00	2.00	6.00
Supplemented		2.00	2.00	2.00	6.00
Pure Panel	0.5	3.69	1.50	0.81	6.00
Supplemented		2.59	1.71	1.16	5.46
Pure Panel	0.7	4.53	1.02	0.45	6.00
Supplemented		2.77	1.35	0.74	4.86
Pure Panel	0.9	5.48	0.38	0.14	6.00
Supplemented		2.93	0.64	0.26	3.83
Pure Panel	0.99	5.95	0.04	0.01	6.00
Supplemented		2.99	0.08	0.03	3.10

The roots of the normalized covariance matrices for the two designs and various values of ρ are given in Table 3. Except for $\rho = 0$, the supplemented panel design has a smaller largest root and a larger smallest root. If one uses the trace criterion to choose a design, then one would always use the design in which some new elements are observed at each time period. For example, if $\rho = 0.5$, the ratio of the trace of the covariance matrix of the pure panel design to that of the supplemented panel design is 1.10. If $\rho = 0.7$ the ratio is 1.23. If one uses the generalized variance criterion, one would always choose a pure panel design. This is due to the fact that the panel design has the smallest small roots.

Table 4. Variances of estimators of parametric functions for five-period supplemented design relative to variances of pure panel design

ρ	μ_1	μ_2	μ_3	$\mu_2 - \mu_1$	$\mu_3 - \mu_1$	$\mu_4 - \mu_1$
−0.70	0.833	0.735	0.721	0.661	1.250	0.627
−0.50	0.928	0.871	0.867	0.765	1.108	0.815
0	1.000	1.000	1.000	1.000	1.000	1.000
0.50	0.928	0.871	0.867	1.301	1.108	1.007
0.70	0.833	0.735	0.721	1.478	1.250	1.104
0.90	0.615	0.517	0.492	1.750	1.577	1.450
0.99	0.375	0.357	0.351	1.967	1.937	1.914

Table 4 contains the variance of some estimators of period means and period-to-period changes for the supplemented panel design relative to the pure panel design for a five-period study under the autoregressive error assumption. In the supplemented panel design, one-sixth of the units are observed in all five periods and each of the remaining five sets of one-sixth of the units is observed in exactly one of the five periods. At each time point, $n/3$ of the units are observed. In the pure panel, the same $n/3$ units are observed at each time point. The variances of the first period means are similar to those of the three-period design. However, the variances of the second and third period means are considerably smaller than the second period mean of the three-period design. The limit of the relative efficiency for the means of the pure panel relative to the supplemented panel as ρ approaches one is one third. On the other hand the limit of the relative efficiency for the difference of two means is 2.0, as it was for the three-period design.

The general structure of the roots of the normalized covariance matrix for the five-period design is similar to that of the three-period design. The largest root of the pure panel design is always larger than the largest root of the supplemented panel design. The supplemented panel design always has a larger smallest root than the pure panel design. The dominance of the supplemented panel relative to the pure panel design with respect to the trace criterion is greater for the five-period design than for the three-period design. This is a further reflection of the fact that estimates of period means are being improved more than estimates of differences are being degraded, as the number of periods is increased. The trace for the five-period pure panel is 15.00 and the traces for the supplemented panel design are 11.57, 13.39, 15.00, 13.39, 11.57, 8.27, and 5.45 for $\rho = -0.70$, -0.50, 0.00, 0.50, 0.70, 0.90, and 0.99, respectively.

The comparisons in Tables 1 through 4 assume a continuous variable and an autoregressive error structure. To compare alternative designs for discrete variables we consider a binomial variable observed at three points in time. A state of zero or one is observed at each time point. It is possible to construct a $2 \times 2 \times 2$ table from such observations. Entries in the table correspond to observations having different states at different time points. There are many alternative ways to construct variables to represent the entries in the $2 \times 2 \times 2$ table. We let

"X_j	$= 1$	if state one at time j , for $j = 1, 2, 3$
" "	$= 0$	otherwise
" "X_{2+j}	$= 1$	if state one at time 1 and state one at time j, for $j = 2, 3$
" "	$= 0$	otherwise
" "X_6	$= 1$	if state one at time 2 and state one at time 3
" "	$= 0$	otherwise
" "X_7	$= 1$	if state one at times 1, 2, and 3
" "	$= 0$	otherwise.
"		

Notice that in the continuous case we defined three variables, but in the binomial case we define seven variables associated with the full $2 \times 2 \times 2$ table.

To define the probability structure of the example we assign probabilities to the two states of zero and one. We assume that the probability of being in state one is 0.5 and the probability of being in state i at time t, given state i at time $t - 1$ is 0.8 for both states zero and one.

The normalized covariance matrices for estimators constructed with the pure panel design and the supplemented panel design are given in Table 5. The entries are for the variance of a sample of size n at each time period multiplied by n. The supplemented panel design has smaller variances for the three time period marginals and for the marginals of the adjacent 2×2 tables (period 1 and 2, and period 2 and 3). The pure panel design has smaller variances for the one-three second order interaction (X_5) and the third order interaction (X_7). For those estimators in the supplemented panel design with smaller variances, the covariances are also smaller than those of the pure panel design. As a result, for example, the estimated change from period one to period two is smaller for the pure panel design than for the supplemented panel design.

The characteristic roots of the normalized covariance matrix for the pure panel design are

$$(1.249, 0.209, 0.113, 0.056, 0.031, 0.010, 0.003)$$

and the characteristic roots for the supplemented panel design are

$$(0.942, 0.265, 0.176, 0.096, 0.059, 0.021, 0.006).$$

Table 5. Normalized covariance matrices for alternative designs for $2 \times 2 \times 2$ table

Covariance Matrix of Pure Panel Design						
0.2500	0.1500	0.0900	0.2000	0.1700	0.1200	0.1600
	0.2500	0.1500	0.2000	0.1500	0.2000	0.1600
		0.2500	0.1200	0.1700	0.2000	0.1600
			0.2400	0.1840	0.1600	0.1920
				0.2244	0.1840	0.2112
					0.2400	0.1920
						0.2176

Covariance Matrix of Supplemented Panel Design						
0.2226	0.0750	0.0274	0.1488	0.1250	0.0512	0.1120
	0.2050	0.0750	0.1400	0.0750	0.1400	0.0928
		0.2226	0.0512	0.1250	0.1488	0.1120
			0.2244	0.1480	0.0956	0.1664
				0.2338	0.1480	0.2144
					0.2244	0.1664
						0.2324

The sum of the roots for the pure panel design is 1.672 and the sum for the supplemented panel design is 1.565. The orthogonal linear combination with largest variance has a variance about 30% smaller for the supplemented panel design. This linear combination is essentially a sum of the seven estimates. The remaining six linear combinations have much smaller variances for both designs, and the variances are smaller for the pure panel design than for the supplemented panel design.

3. The National Resources Inventory

The National Resources Inventory is a survey of the nonfederal land area of the United States conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. It is a large survey of about 300,000 primary sampling units, where a primary sampling unit is a segment of land. In the Midwest the segment is 160 acres, but the sizes of the primary sampling units vary across the country. In the western part of the U.S., there are some that are as big as 640 acres and in the East some segments are on the order of 100 acres. Within the primary sampling unit, points are designated for observation. There are either two or three points per primary sampling unit. The basic survey has been conducted in each of the years 1982, 1987, 1992, and processing is currently underway for 1997 data.

In 1995, a sample of 3,000 segments was selected from the 300,000 for a study that was called the Erosion Update Study.This study was a subsample of the large sample, but the primary sampling units were different.In 26 states, counties were sampling units and for 22 of the states, states were primary sampling units. The 1992 basic NRI sample of 300,000 segments and the 1995 subsample of 3,000 segments form a classical two-phase sample. In the first-phase sample of 300,000 sampling units, the vector X is observed. In the subsample of 3,000 units, an extended vector (X, Y) is observed.

A second special study was conducted in 1996. The original 1995 sample of 3,000 was augmented by another 1,000 segments to obtain a total of 4,000 segments. A third study was conducted in 1997 in which the 1996 sample was augmented by another 2,000 segments. The designs for 1996 and 1997 were extensions of the 1995 design with states and counties as primary sampling units. We can divide the original 300,000 segments into four subgroups: 294,000 are observed only in 1992; 3,000 are observed in 1992, 1995, 1996, and 1997; 1,000 are observed in 1992, 1996, and 1997; and 2,000 are observed in 1992 and 1997.

The regression model for a characteristic y for these data is

$$\begin{bmatrix} \overline{y}_{92,294} \\ \overline{y}_{92,3} \\ \overline{y}_{95,3} \\ \overline{y}_{96,3} \\ \overline{y}_{97,3} \\ \overline{y}_{92,1} \\ \overline{y}_{96,1} \\ \overline{y}_{97,1} \\ \overline{y}_{92,2} \\ \overline{y}_{97,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{92} \\ \mu_{95} \\ \mu_{96} \\ \mu_{97} \end{bmatrix} + \begin{bmatrix} e_{92,294} \\ e_{92,3} \\ e_{95,3} \\ e_{96,3} \\ e_{97,3} \\ e_{92,1} \\ e_{96,1} \\ e_{97,1} \\ e_{92,2} \\ e_{97,2} \end{bmatrix}.$$

where $\overline{y}_{t,j}$ is the mean for the characteristic at time t on a group of $j \times (1000)$ segments, and μ_t is the mean of y at time t. The estimation method used in the actual study was closely related to the generalized least squares procedure. Estimates were constructed for a selected set of variables. These variables were then used as controls in a regression type weighting procedure for the general survey. (See Fuller (1990)).

The covariance structure for several characteristics observed in the National Resources Inventory have been estimated. (See Breidt and Fuller (1998)). We use those correlations to study the efficiency of generalized least squares. The efficiencies presented under the model assumptions are approximations because the effect of the clustering of observations into large primary sampling units is largely ignored. Under the model used to estimate the correlations, the observation on segment i at time t, denoted by y_{it}, is expressed as

$$y_{it} = \mu_t + e_{it},$$

where μ_t is the mean at time t and e_{it} is the deviation from the mean. The deviation is assumed to satisfy the model

$$e_{it} = \alpha_{it} + \epsilon_{it},$$

$$\alpha_{it} = \rho\alpha_{i,t-1} + u_{it},$$

where α_{it} is a stationary autoregressive process, $|\rho| < 1$ is the autoregressive parameter, ϵ_{it} is uncorrelated measurement error distributed with mean zero and variance σ^2_ϵ, denoted by $(0, \sigma^2_\epsilon)$, and u_{it} is the $(0, \sigma^2_u)$ uncorrelated error of the autoregressive process. Under the model

$$V\{y_{it}\} = (1 - \rho^2)^{-1}\sigma^2_u + \sigma^2_\epsilon.$$

We let

$$\nu = [V\{y_{it}\}]^{-1}\sigma^2_\epsilon,$$

where ν is the fraction of total variance that is due to measurement error. Under the model the correlation between observations h periods apart is

$$Corr(y_{it}, y_{i,t+h}) = \rho^{|h|}(1 - v), \quad h = \pm 1, \pm 2, \ldots.$$

The estimated parameters for several characteristics are given in Table 6.

Table 6. Estimated parameters of error model

	Parameter	
Characteristic	ρ	ν
USLE	0.955	0.229
Cultivated cropland	0.981	0.014
Noncult. cropland	0.955	0.130
Pasture	0.974	0.031
Rangeland	0.992	0.011
Forest	0.981	0.015
Small built-up	0.957	0.029
Large urban	0.977	0.016
Streams	0.996	0.089
Roads	0.998	0.027

The USLE is the tons of soil loss calculated by the Universal Soil Loss Equation. The computations require a number of factors such as the slope of the land and the cropping practices. Hence, the variable has large relative measurement error with nearly 23% of total variance attributable to measurement error. The remaining variables are the acres in that land use in the segment. Land uses that are relatively permanent, such as rangeland, streams and roads, have large autoregressive coefficients. There is considerable measurement error associated with the determination of the acres in streams.

One simple estimator of the mean for each period is the mean of segments actually observed during that period. Table 7 contains estimated efficiencies of generalized least squares relative to the simple mean. The generalized least squares estimator for 1992 is the simple mean because all other samples are subsamples of the 1992 sample. The estimator for 1997 is a two-phase estimator with the 1992 sample as the first phase sample. The estimates for 1995 and 1996 use information from more than one other sample.

Table 7. Estimated relative efficiency of generalized least squares to simple mean

	Year			
Characteristic	1992	1995	1996	1997
USLE	1.00	2.12	1.88	1.58
Cultivated cropland	1.00	9.53	7.13	4.67
Noncult. cropland	1.00	2.87	2.43	1.89
Pasture	1.00	6.47	5.01	3.43
Rangeland	1.00	16.66	12.73	8.59
Forest	1.00	9.59	7.18	4.71
Small built-up	1.00	4.79	3.67	2.47
Large urban	1.00	8.26	6.18	4.05
Streams	1.00	5.97	5.38	4.65
Roads	1.00	16.08	13.84	11.12

The smallest efficiency gains are for USLE which has the smallest autocorrelations. The year-to-year correlation for USLE is about 0.74. On the other hand, generalized least squares is much superior to the simple mean for rangeland which has a year-to-year correlation of about 0.98. The estimated efficiencies for changes are given in Table 8. The gains for changes are generally comparable to those for levels.

Table 8. Estimated relative efficiency of generalized least squares to simple means for changes

	Change		
Characteristic	$\mu_{92} - \mu_{95}$	$\mu_{95} - \mu_{96}$	$\mu_{96} - \mu_{97}$
USLE	2.13	1.33	1.41
Cultivated cropland	10.29	4.96	6.09
Noncult. cropland	2.90	1.62	1.79
Pasture	6.76	3.24	3.88
Rangeland	19.53	7.83	9.87
Forest	10.36	4.96	6.10
Small built-up	4.92	2.75	3.22
Large urban	8.80	4.40	5.37
Streams	6.22	2.30	2.70
Roads	18.75	5.57	7.03

A large portion of the gains exhibited in Tables 7 and 8 comes from the large sample observed in 1992. It would be possible to estimate the mean for each year treating the sample for that year as the second phase of a two-phase sample where the first phase is the 1992 sample. Table 9 contains the efficiencies of the generalized least squares estimates of the year means relative to the two-phase estimator. The efficiency of the two-phase estimator is 1.00 for 1997 because the 1995 and 1996 samples are subsamples of the 1997 sample. The generalized least squares estimator for 1996 uses information from 1992 and 1997 and the generalized least squares estimator for 1995 uses information from 1992, 1997, and 1996. The use of all information produces gains on the order of 20% to 30% for 1995 relative to the use of only 1992 information, for the characteristics investigated. The gains for 1996 from adding the 1997 information are about 60% to 75% of the gains for 1995.

Table 9. ESTIMATED RELATIVE EFFICIENCY OF GENERALIZED LEAST SQUARES TO TWO-PHASE ESTIMATOR

	Parameter		
Characteristic	μ_{95}	μ_{96}	μ_{97}
USLE	1.17	1.12	1.00
Cultivated cropland	1.36	1.27	1.00
Noncult. cropland	1.23	1.17	1.00
Pasture	1.33	1.25	1.00
Rangeland	1.30	1.22	1.00
Forest	1.36	1.27	1.00
Small built-up	1.36	1.27	1.00
Large urban	1.37	1.27	1.00
Streams	1.17	1.10	1.00
Roads	1.18	1.11	1.00

4. **Comments**

The properties of supplemented panel designs combined with generalized least squares estimation were investigated for surveys conducted at multiple time points. Supplemented panels are appealing designs when respondent burden is not a serious consideration. The panel component is efficient for period-to-period changes and the supplemented samples improve the efficiency of estimates of level. Gains are largest when correlations are high and gains in efficiency for levels increase as the length of the study increases. The estimated correlations for the National Resources Inventory demonstrate that the supplemental panel generalized least squares procedures can produce large gains over simple mean estimators.

References

BREAU, P. AND ERNST, L. R. (1983). Alternative estimators to the current composite estimator. *Proc. of the Section on Survey Research Methods*, Amer. Statist. Assoc., 397-402.

BREIDT, F. J. AND FULLER, W. A. (1998). Design of supplemented panel surveys with application to the National Resources Inventory. *Jour. Agricultural, Biological, and Ecological Statistics*, to appear.

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Jour. Amer. Statist. Assoc.*, **37**, 199-212.
Duncan, G. J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, **55**, 97-117.
Eckler, A. R. (1955). Rotation sampling. *Ann. Math. Statist.*, **26**, 664-685.
Fuller, W. A. (1990). Analysis of repeated surveys. *Survey Methodology*, **16**, 167-180.
–––– (1998). Replication variance estimation for two phase samples. *Statistica Sinica*, **8**, 1153-1164.
Gurney, M. and Daly, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proc. of the American Statistical Association, Section on Social Statistics*, 242-257.
Huang, L. R. and Ernst, L. R. (1981). Comparison of an alternative estimator to the current composite estimator in the Current Population Survey. *Proc. of the Amer. Statist. Assoc., Section on Survey Research Methods*, 303-308.
Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, **304**, 54-59.
Jones, R. G. (1980). Best linear unbiased estimators for repeated surveys. *Jour. Royal Statist. Soc.*, Series B, **42**, 221-226.
Kasprzyk, D., Duncan, G. J., Kalton, G., and Singh, M. P. (1989). *Panel Surveys*, Wiley, New York.
Kish, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, **28**, 1-12.
Kumar, S. and Lee, H. (1983). Evaluation of composite estimation for the Canadian Labor Force Survey. *Survey Methodology*, **9**, 1-24.
Lent, J., Miller, S. M., and Cantwell, P. J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proc. of the Section on Survey Research Methods*, Amer. Statist. Assoc., 130-139.
Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Jour. Royal Statist. Soc.*, Series B, **12**, 241-255.
Raj, D. (1965). On sampling over two occasions with probability proportionate to size. *Ann. Math. Statist.*, **36**, 327-330.
Rao, J. N. K. and Graham, J. E. (1964). Rotation designs for sampling on repeated occasions. *Jour. Amer. Statist. Assoc.*, **59**, 492-509.
Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453-460.
Schreuder, H. T., Gregoire, T. G., and Wood, G. B. (1993). *Sampling Methods for Multi-Resource Forest Inventory.* Wiley, New York.
Singh, A. C. (1996). Combining information in survey sampling by modified regression. *Proc. of the Section on Survey Research Methods*, American Statistical Association, 120-129.
Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Jour. Amer. Statist. Assoc.*, **92**, 780-787.
Smith, T. M. F. (1978). Principles and problems in the analysis of repeated surveys. Pages 201-216. In N. Krishnan Namboodiri, ed. *Survey Sampling and Measurement.* Academic Press, New York.
Wolter, K. (1979). Composite estimation in finite populations. *Jour. Amer. Statist. Assoc.*, **74**, 604-613.
Yansaneh, I. S. and Fuller, W. A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*, **24**, 31-40.

Wayne A. Fuller
Department of Statistics
Iowa State University
218 Snedecor Hall
Ames, IA 50011-1210
e-mail : waf@iastate.edu

F. Jay Breidt
Department of Statistics
Iowa State University
221 Snedecor Hall
Ames, IA 50011-1210
e-mail : jbreidt@iastate.edu