Survey Methodology, June 1998
Vol. 24, No. 1, pp. 31-40
Statistics Canada

Optimal Recursive Estimation for Repeated Surveys

IBRAHIM S. YANSANEH and WAYNE A. FULLER[1]

ABSTRACT

Least squares estimation for repeated surveys is addressed. Several estimators of current level, change in level and average level for multiple time periods are developed. The Recursive Regression Estimator, a recursive computational form of the best linear unbiased estimator based on all periods of the survey, is presented. It is shown that the recursive regression procedure converges; and that the dimension of the estimation problem is bounded as the number of periods increases indefinitely. The recursive procedure offers a solution to the problem of computational complexity associated with minimum variance unbiased estimation in repeated surveys. Data from the U.S. Current Population Survey are used to compare alternative estimators under two types of rotation designs: the intermittent rotation design used in the U.S. Current Population Survey, and two continuous rotation designs.

KEY WORDS: Recursive regression estimation; Composite estimation; Rotation designs; Rotation groups.

1. INTRODUCTION

We consider least squares estimation for surveys conducted on repeated occasions with partial overlap of sampling units. See Duncan and Kalton (1987) for a general discussion of different types of surveys and the objectives of such surveys. In this paper, we shall be concerned with rotating panel surveys, in which repeated determinations are made on some sampling units but not every unit appears in the sample at every time point.

Theoretical foundations for the design and estimation for repeated surveys based on generalized least squares procedures were laid down by Patterson (1950), following initial work by Cochran (1942) and Jessen (1942). Least squares procedures have been considered further by several other authors. See, for instance, Fuller (1990), and the references cited therein. Least squares estimation for a fairly general class of repeated surveys was considered by Yansaneh (1992). Composite estimation is a procedure of estimation for repeated surveys which makes use of the observations from the current and preceding periods, and the estimator of level from the preceding period. Breau and Ernst (1983) compared various alternative estimators to a composite estimator for the U.S. Current Population Survey (CPS). Kumar and Lee (1983) did similar work using data from the Canadian Labor Force Survey (LFS). Wolter (1979) provided a general composite estimation strategy for two-level rotation schemes such as the one used in the U.S. Census Bureau's Retail Trade Survey. Singh (1996) has proposed an alternative class of composite estimators. These authors assumed the unknown quantities on each occasion to be fixed parameters. Other authors, such as Scott, Smith, and Jones (1977), Jones (1980), Binder and Dick (1989), Bell and Hillmer (1990), and Pfeffermann (1991) considered estimation for repeated surveys under the assumption that the underlying true values constitute a realization of a time series.

In this paper, we discuss estimation procedures for repeated surveys, under the assumption that the unknown true values are fixed parameters. The estimators are compared to the method of composite estimation currently used in the CPS. The paper is organized as follows: In section 2, we state some basic assumptions regarding the general class of repeated surveys considered in this paper. A description of the CPS method of composite estimation is given in section 3. The method of best linear unbiased estimation is discussed in section 4. In section 5, we present a recursive regression estimation procedure designed to reduce the computational complexity associated with best linear unbiased estimation. Section 6 is devoted to an application to data from the CPS. Alternative estimators and rotation designs are compared.

2. BASIC ASSUMPTIONS

In this section, we describe surveys of the type we will study. A rotation group is a set of individuals selected for the sample and observed for a fixed number of periods and in a fixed pattern over time. Assume that in each period of the survey, s rotation groups are included in the sample, where $s > 1$ is fixed. Assume that the basic data from the survey can be organized in a set of elementary estimators (such as simple sample means and estimated totals) of the parameters of interest (such as population means and totals), where a set of elementary estimators is associated with each rotation group. For computational convenience, the data for p periods can be arranged in a $p \times s$ data matrix, denoted by $\boldsymbol{H}$, in such a way that the observations on a rotation group appear in only one column. The total

[1] Ibrahim S. Yansaneh, Statistical Group, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850; and Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50011 U.S.A.

number of elementary estimators is $n = p \times s$. We call the columns of $\boldsymbol{H}$ streams. Several rotation groups can appear in a stream. Assume that:

(1) A given rotation group in a stream is observed over a period of total length $m+1$, and the observation pattern for rotation groups is fixed and is the same for all groups.

(2) The design is balanced on time-in-sample. That is, of the s rotation groups included in the sample at a given time, one group is being observed for the first time, one is being observed for the second time, ..., one is being observed for the last time, where the last time is separated by m periods from the first observation.

These assumptions are satisfied by surveys such as the CPS and the Canadian Labor Force Survey. See Yansaneh (1997) for an illustration of the 4-8-4 rotation scheme used in the CPS.

3. THE CPS COMPOSITE ESTIMATOR

In general, composite estimators combine recent estimator(s) and data from the current and preceding period(s) to form an estimator for the current period. With the CPS, six of the eight rotation groups observed at time t were observed at time $t-1$. We shall refer to these six rotation groups as continuing rotation groups, and the remaining two as incoming rotation groups.

The composite estimator currently in use is determined by two parameters. The estimator is

$$\hat{\theta}_{t,c} = (1-\pi_1)\bar{y}_t + \pi_1(\hat{\theta}_{t-1,c} + \hat{\delta}_{t,t-1}) + \pi_2\hat{\delta}_t \qquad (1)$$

where, for the estimator currently used, $\pi_1 = 0.4$ and $\pi_2 = 0.2$, $y_{t,k}$ is the elementary estimate of level obtained from the rotation group which is in its k-th time in sample at time t, $\bar{y}_t = 8^{-1}\sum_{k=1}^{8} y_{t,k}$ is the basic estimator, defined as the mean of the elementary estimates based on the eight rotation groups observed at time t, $\hat{\theta}_{t-1,c}$ is the composite estimator for time $t-1$, $\hat{\delta}_{t,t-1}$ is an estimate of change in level, based on the six continuing rotation groups at time t, and $\hat{\delta}_t$ is the difference between the averages of the two incoming rotation groups and the six continuing rotation groups. Thus,

$$\hat{\delta}_{t,t-1} = 6^{-1}\sum_{k\in S}\left(y_{t,k} - y_{t-1,k-1}\right),$$

and

$$\hat{\delta}_t = 8^{-1}\left(\sum_{k\in T} y_{t,k} - 3^{-1}\sum_{k\in S} y_{t,k}\right),$$

where $T = \{1,5\}$ and $S = \{2,3,4,6,7,8\}$. The composite estimator used until 1985 contained only the first two terms on the right of (1). The third term was introduced for the purpose of reducing the time-in-sample effects appearing in the original estimator. The incoming rotation groups produce larger estimates of unemployed than do the continuing rotation groups. Therefore, the direct difference $\hat{\delta}_{t,t-1}$ is influenced by the fact that the rotation group in its first time-in-sample has a larger expected value than that of the second time-in-sample. The time-in-sample effects do not cancel out in the difference estimate. The third term is an adjustment term which has the effect of reducing both the variance of the original composite estimator and the bias associated with time-in-sample effects. See Bailar (1975) or Breau and Ernst (1983) for a discussion of the bias of the pre-1985 composite estimator due to time-in-sample effects. We shall refer to the three-term composite estimator currently used in the CPS as the CPS Composite Estimator. This estimator has a variance close to that of the best linear unbiased estimator for monthly estimates of unemployment level. Let $y_{i,t}$, $i = 1, 2, ..., s$, be the elementary estimator of the parameter of interest obtained from the rotation group which is in stream i at time t. The CPS composite estimator can be written as

$$\hat{\theta}_{t,c} = \sum_{i=1}^{8}\omega_{1,k(i,t)}y_{i,t} + \sum_{i=1}^{8}\omega_{2,k(i,t)}y_{i,t-1} + \pi_1\hat{\theta}_{t-1,c} \qquad (2)$$

where $k(i,t) = k$ defines the time-in-sample of observation (it) as a function of the stream (i) and time (t). If $\lambda_1 = 1/8$ and $\lambda_2 = -1/6$, and $\lambda_3 = 1/3$, then $\omega_{2,k} = \pi_1\lambda_2$, and

$$\omega_{1,k} = \begin{cases}(1-\pi_1)\lambda_2 - \pi_1\lambda_2 - \pi_2\lambda_1\lambda_3 & \text{for } k\in S\\ \lambda_1(1-\pi_1+\pi_2) & \text{for } k\in T\end{cases}$$

Let

$$\boldsymbol{p}_1 = \left(\omega_{1,k(1,t)}, \omega_{1,k(2,t)}, ..., \omega_{1,k(8,t)}\right)',$$

$$\boldsymbol{p}_2 = \left(\omega_{2,k(1,t)}, \omega_{2,k(2,t)}, ..., \omega_{2,k(8,t)}\right)',$$

and $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, ..., y_{8,t})'$. Then,

$$\hat{\theta}_{t,c} = \boldsymbol{p}_i'\boldsymbol{y}_t + \boldsymbol{p}_2'\boldsymbol{y}_{t-1} + \pi_1\hat{\theta}_{t-1,c} \qquad (3)$$

Substituting in (3) recursively, we have, for an estimator initiated at time zero,

$$\hat{\theta}_{t,c} = \boldsymbol{p}_1'\boldsymbol{y}_t + \sum_{j=1}^{t}\pi_1^{j-1}(\boldsymbol{p}_2 + \pi_1\boldsymbol{p}_1)'\boldsymbol{y}_{t-1} \qquad (4)$$

Equation (4) is an expression of $\hat{\theta}_{t,c}$ as a linear function of current and past observations, where the weight of an observation declines as its distance from the current period increases.

4. BEST LINEAR UNBIASED ESTIMATION

Suppose $\Theta_p = (\theta_1, \theta_2, ..., \theta_p)'$ is the $p \times 1$ vector of parameters of interest, where θ_t, $t = 1, 2, ..., p$, is the level of the parameter of interest at time t. Thus at time j, θ_j is the current level of the parameter of interest. For example, in the context of the CPS, θ_j might represent the population mean or proportion of unemployed at time j. Our objective is to construct efficient estimators of the current level of the parameters. The change in level and average level over multiple periods of time are also of interest.

The best linear unbiased estimator (BLUE) of the current level is defined to be the minimum-variance unbiased linear combination of the elementary estimators from the rotation groups available for estimation. It is possible in the process of computing the BLUE for the current level, to also compute the BLUEs for all periods using data available at the current time.

Suppose that a repeated survey has been in operation for p periods and that s streams of data collected over p periods are available for estimation. Let $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, ..., y_{i,p})'$ be the vector of p observations in the i-th stream at time t. Let $\boldsymbol{Y}_p$ be the data vector formed by the streams or columns of the $p \times s$ data matrix $\boldsymbol{H}$, arranged chronologically. Thus, $\boldsymbol{Y}_p = (\boldsymbol{y}_1', \boldsymbol{y}_2', ..., \boldsymbol{y}_s')'$ is an $n \times 1$ vector of observations, where $n = s \times p$. Let $\boldsymbol{X} = J_{s\times 1} \otimes I_{p\times p}$ be the $n \times p$ design matrix which relates the estimates in $\boldsymbol{Y}_p$ to their expected values in Θ_p; where $J_{s\times 1}$ is the $s \times 1$ vector of ones, $I_{p\times p}$ is the identity matrix of order p, and $\otimes$ denotes the Kronecker product. The linear model for $\boldsymbol{Y}_p$ is

$$\boldsymbol{Y}_p = \boldsymbol{X}\Theta_p + \boldsymbol{U}_p \tag{5}$$

where $\boldsymbol{U}_p$ is the vector of error terms satisfying the assumptions $\mathrm{E}(\boldsymbol{U}_p) = 0$ and $\mathrm{E}(\boldsymbol{U}_p\boldsymbol{U}_p') = \boldsymbol{V}_p$, where $\boldsymbol{V}_p$ is assumed to be a known, symmetric, and positive definite matrix. By the Gauss-Markov Theorem, the BLUE of Θ_p is

$$\hat{\Theta}_p = (\boldsymbol{X}'\boldsymbol{V}_p^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_p^{-1}\boldsymbol{Y}_p.$$

The covariance matrix of $\hat{\Theta}_p$ is $\sum_p = (\boldsymbol{X}'\boldsymbol{V}_p^{-1}\boldsymbol{X})^{-1}$.

5. RECURSIVE REGRESSION ESTIMATION

Recursive estimation techniques have been found useful in situations where data do not all become available at the same time but rather accumulate over time, and the computation of optimal estimates based on all available data is impractical. See, for example, Odell and Lewis (1971), Sallas and Harville (1981) and references cited therein, for recursive algorithms for best linear unbiased estimation. Tiller (1989) presented a Kalman-filter approach to estimation of labor force characteristics using survey data.

As described in Section 4, the direct computation of the BLUE becomes progressively more complicated as the number of periods increases. We develop a recursive regression estimation procedure for repeated surveys that uses a judiciously chosen set of initial estimates, new observations of the current level, and the previous observations on the currently observed rotation groups to produce the BLUE of current level.

5.1 Transformed Elementary Estimates and a Proposed Estimator

Suppose a survey has been in operation for at least m periods and assume:

(3) The rotation groups are independent.

(4) The covariance structure of the observations is known.

(5) The covariance structure of the observations in a stream is constant over time, and it is the same for all streams.

These assumptions are used in the construction of a linear estimator. Assumption (3) will be relaxed for the computation of the variance of the estimator. Under assumptions (1) and (3), observations that are more than m periods apart are independent. At the current time, denoted by c, where $c > m$, a set of s elementary estimates of the parameter θ_c are observed. To construct the generalized least squares estimator, the s current observations are transformed so that they are uncorrelated with previous observations. After transformation, the expected values of the transformed observations are functions of θ_c and the parameters for the m preceding periods. Assume that the BLUE of the vector of parameters for the previous m periods, and the $m \times m$ covariance matrix of these estimators, are available. Thus, at time c, we have: (i) m initial estimates $\hat{\Theta}_{c-1(m)} = (\hat{\theta}_{c-m}, ..., \hat{\theta}_{c-1})'$; (ii) the covariance matrix $\sum_{c-1(m)}$ of $\hat{\Theta}_{c-1(m)}$; and (iii) s independent observations on the s streams at the current time. Let the transformed observations, denoted by z_{ic}, $i = 1, 2, ..., s$, be

$$z_{ic} = y_{i,c} - \sum_{j=1}^{m} b_{k(i,c),j}\, y_{i,c-j} \tag{6}$$

where $b_{k(i,c),j}$ are the coefficients such that $z_{i,c}$ is uncorrelated with $y_{i,c-j}$ for all $j > 0$. By assumptions (4) and (5), the coefficients $b_{k(i,c),j}$ are fixed over time. By assumption (3), $z_{i,c}$ is uncorrelated with all earlier observations. The expected value of $z_{i,c}$ is $\theta_c - \sum_{j=1}^{m} b_{k(i,c),j}\theta_{c-j}$, $i = 1, 2, ..., s$.

5.2 The Recursive Regression Estimator

Let $\hat{\theta}_h(t)$, $h \le t$, denote the least squares estimator of the (scalar) parameter θ_h constructed using data through time t; and let $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), ..., \hat{\theta}_t(t))'$ denote the least squares estimator of the vector of m parameters $\theta_{t-m+1}, ..., \theta_t$, at time t constructed using data through time t. Our objective is to construct the minimum variance estimator for θ_c, the current level of the parameter of interest using all data available at time c. A linear model for data available at the current time is

$$Z_c = W\Theta_{c(m+1)} + U_c \tag{7}$$

where

$$W = \begin{pmatrix} I_m & 0 \\ X_{21} & J_s \end{pmatrix},$$

$Z_c' = (\hat{\Theta}_{c-1(m)}', z_c'), z_c' = (z_{1c}, ..., z_{sc})$, and X_{21} is an $s \times m$ matrix whose entries are constant over time, and are functions of the coefficients $b_{k,j}$ of (6). If $\mathrm{Var}\{z_{i,c}\} = \sigma_i^2$, $i = 1, 2, ..., s$, and Ω_{22} is the diagonal matrix with σ_i^2 as the diagonal entries, then the covariance matrix of Z_c is $V_c = \mathrm{blockdiag}\{\sum_{c-1(m)}, \Omega_{22}\}$. It is assumed that σ_i^2, $i = 1, 2, ..., s$, are positive.

The recursive regression estimator (RRE) of $\Theta_{c(m+1)}$ is defined to be the least squares estimator of $\Theta_{c(m+1)}$ based on model (7). Thus the RRE of $\Theta_{c(m+1)}$ is

$$\hat{\Theta}_{c(m+1)} = (W' V_c^{-1} W)^{-1} W' V_c^{-1} Z_c \tag{8}$$

and the covariance matrix of $\hat{\Theta}_{c(m+1)}$ is $Q_{c(m+1)} = (W' V_c^{-1} W)^{-1}$.

The utility of the estimator (8) is its computational simplicity. At any fixed time t in a repeated survey, all the information relevant to the problem of estimating $\theta_t, \theta_{t-1}, ..., \theta_{t-m}$ can be obtained from a set of m recursive least squares estimates and the current observations.

We now describe more fully the recursive regression procedure. At time t, we have $\hat{\Theta}_{t(m+1)}$, the RRE of $\Theta_{t(m+1)}$, and its $(m+1) \times (m+1)$ covariance matrix $\sum_{t(m+1)}$. Partition $\sum_{t(m+1)}$ as

$$\sum\nolimits_{t(m+1)} = \begin{pmatrix} v_{11,t} & V_{12,t} \\ V_{12,t}' & \sum_{t(m)} \end{pmatrix},$$

where $v_{11,t}$ is the variance of $\hat{\theta}_{t-m}(t), \sum_{t(m)}$ is the covariance matrix of $(\hat{\theta}_{t-m+1}(t), ..., \hat{\theta}_t(t))'$, and $V_{12,t}$ is the covariance between these two quantities. Observe that if θ_{t-m} is retained in the parameter vector and $\hat{\theta}_{t-m}$ is retained in the data vector, the estimator of θ_{t+1} is unchanged (the estimator of θ_{t-m} would, in general, be changed). This is because the estimator of the original parameter vector of a least squares problem is not changed if an observation whose expectation is equal to a single new parameter is added to the problem. Thus, to update the RRE for the next period, we drop the initial estimate for the earliest period, $\hat{\theta}_{t-m}(t)$, from the data vector, and drop the corresponding parameter θ_{t-m} from the parameter vector. The parameter θ_{t+1} is then added to the parameter vector. In this way, the dimension of the basic model matrix W of the estimation problem is kept constant over time. Thus in the class of repeated surveys considered in this paper, there is an upper bound on the computational effort required for the BLUE of the vector of parameters of interest.

The model at time $t+1$ may be written as model (7), with $c = t+1, Z_{t+1} = (\hat{\theta}_{t-m+1}(t), ..., \hat{\theta}_{t-1}(t), \hat{\theta}_t(t), z_{t+1}')'$, $\Theta_{t+1(m+1)} = (\theta_{t-m+1}, ..., \theta_t, \theta_{t+1})'$, and the covariance matrix of Z_{t+1} is $V_{t+1} = \mathrm{blockdiag}\{\sum_{t(m)}, \Omega_{22}\}$. The BLUE of $\Theta_{t+1(m+1)}$ and its covariance matrix are then obtained from the usual least squares formulas. The least squares estimators of the last m elements of $\Theta_{t+1(m+1)}$ are then used as the initial estimates in the model for the next iteration.

The following theorem states that the covariance matrix of the vector of recursive least squares estimators converges to a positive definite matrix as the number of periods in the survey increases indefinitely. A proof is given in the appendix.

Theorem: At any time t, let the vector of recursive least squares estimators $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), ..., \hat{\theta}_{t-1}(t), \hat{\theta}_t(t))'$ be the BLUE of the vector of parameters $\Theta_{t(m)} = (\theta_{t-m+1}, ..., \theta_{t-1}, \theta_t)'$ based on data through time t. Let $\sum_{t(m)}$ be the covariance matrix of $\hat{\Theta}_{t(m)}$. Let the assumptions (1) through (5) hold. Also assume that the elements of V_n^{-1} are bounded for all n, where V_n is the covariance matrix of any n observations. Then, the covariance matrix $\sum_{t(m)}$ converges as $t \to \infty$; and the limit is an $m \times m$ positive definite matrix.

6. APPLICATION TO THE U.S. CURRENT POPULATION SURVEY

6.1 The CPS Design

The CPS is a monthly household survey conducted by the United States Census Bureau in cooperation with the Bureau of Labor Statistics for the purpose of providing national estimates of labor force characteristics such as the number employed, unemployed, and in the civilian labor force; and other characteristics of the non-institutionalized civilian population. The sample design of the CPS contains a rotation scheme that includes the replacement of a fraction of the households in the sample each month. For any given month, the sample consists of eight time-in-sample panels or rotation groups, of which one is being interviewed for the first time, one is being interviewed for the second time,..., and one is being interviewed for the eighth time. In other words, the interview scheme is balanced on time-in-sample. Households in a rotation group are interviewed for four consecutive months, dropped for the next eight succeeding months, and then interviewed for another four consecutive months. They are then dropped from the sample entirely. This system of interviewing is called the 4-8-4 rotation scheme, and is a special case of schemes described by Rao and Graham (1964).

6.2 Estimation and Variance Estimation Procedures

We use estimates of the covariance structure of data from the CPS to compare alternative estimators and rotation designs. See Adam and Fuller (1992) and Fuller, Adam and Yansaneh (1993) for a detailed description of the construction of the model, the estimation of its parameters, and the estimation of the covariance structure of observations within a given rotation group for various characteristics of interest. Because the rotation groups come from the same set of primary sampling units, they are not independent and a component is included in the covariances to reflect the fact that the primary sampling units do not change. The RRE is computed with the eight current simple estimators and the 15 estimators for the 15 preceding periods. In computing the RRE, the covariances are used to create eight linear combinations of the current and the preceding fifteen observations that are uncorrelated with the preceding fifteen observations. Because of the primary sampling unit effect, these linear combinations are correlated with observations more than 15 periods in the past and in the same stream. Hence, they are correlated with the preceding estimators. The correlations with earlier estimators, $\hat{\theta}_{t-i}$, $i = 1, ..., 15$, are included in the covariance matrix when the estimator of θ_t is constructed. However, because only the most recent 15 observations are used, the resultant estimator of θ_t is not the BLUE for current level. The calculated covariance matrix of $(\hat{\theta}_{t-15}, ..., \hat{\theta}_{t-1}, \hat{\theta}_t)'$ is correct and, because the primary sampling unit effect is modest, it is felt that the estimator has efficiency close to that of the BLUE.

We shall restrict attention to the estimation of various parameters for two characteristics of interest: Employed and Unemployed. For each characteristic, the parameters of interest are the current level and period-to-period change for up to 12 periods. The estimators considered for comparison are the CPS composite estimator; the RRE; and the BLUEs using 2, 3, 12, 16, and 24 periods, where the BLUE for p periods at time t is the least squares estimator constructed using data from time $t - p + 1$ through time t. Results are reported for BLUEs based on 12 and 16 periods. In following the practice of the U.S. Bureau of Labor Statistics for CPS estimators, the estimators are not modified as new data become available. Thus the estimator of change in level of a characteristic of interest between times $t - 1$ and t is not the best possible estimator given all available data. It is the difference between the best estimator at time t based on data through time t and the best estimator at time $t - 1$ based on data through time $t - 1$.

We do not consider seasonal adjustment in this discussion. However, the estimation procedures presented can be extended to include seasonal adjustments. To compute the variance of a given estimator at a given time, the estimator is first expressed as a linear combination of all the observations available at that time. The variance of the estimator is then computed as a function of the coefficients of the linear combination and the entries of the covariance matrix.

6.3 Numerical Results and Discussion

6.3.1 Comparison of Alternative Estimators

The variances of the alternative estimators relative to the variance of the basic estimator of current level, for each of the characteristics of interest, are presented in Table 1. Recall that the basic estimator of the current level, denoted by $\bar{y}_t$ is the simple mean of the eight elementary estimators obtained from the eight rotation groups observed at time t. That is, $\bar{y}_t = 8^{-1}\sum_{k=1}^{8} y_{t,k}$, and $\mathrm{Var}(\bar{y}_t) = \sigma^2/8$, where $\sigma^2 = \mathrm{Var}(y_{t,k})$ for all t and k. The basic estimator of change between two periods is the difference between the simple means for the two periods.

The BLUE procedure based on 3 periods or more produces more efficient estimators of current level than the CPS composite estimator. In general, the best linear unbiased estimation procedure becomes more statistically efficient as the number of periods increases. For both characteristics, the results reveal that the best linear unbiased procedure based on 12 periods is uniformly more efficient than the CPS composite estimator for all parameters, except one-period change in unemployed. Recall that the estimator of change is not BLUE because the estimator is the difference of estimators constructed at time t and at time $t - 1$. Thus, the estimator called "BLUE" is best only for current level using the stated amount of data. The difference between the variance of the composite estimator of one-period change and the variance of the 12-period BLUE of one-period change in unemployed is less than one percent. The gain in precision of the best linear unbiased estimation procedure for employed relative to the CPS composite estimator for current level is 22% for the BLUE for 12 periods, 28% for the BLUE for 16 periods, 30% for the BLUE for 24 periods, and 33% for the RRE. The corresponding gains for unemployed are 2%, 3%, and 3%. These results are a reflection of the nature of the autocorrelation functions of the characteristics. The autocorrelation function for unemployed declines much faster than that for employed.

With the exception of one-period change in employed, there is an improvement in the efficiency of the estimation of change from using the alternative estimators instead of the CPS composite estimator. The gain in precision increases as the number of periods increases, reaching a maximum value at five-period change for both characteristics. The gain then decreases slightly. In the case of the RRE, the maximum gain in efficiency for estimated change is 64% for employed and 5% for unemployed.

Table 1
Variances of alternative estimators relative to the variance of the basic estimator of current level

	Employed				Unemployed			
Parameter	CPS Comp.	BLUE for 12 periods	BLUE for 16 periods	Recursive Regression Estimator	CPS Comp.	BLUE for 12 Periods	BLUE for 16 periods	Recursive Regression Estimator
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.704	0.672	0.650	0.947	0.924	0.918	0.918
1-period change	0.511	0.457	0.437	0.432	1.070	1.077	1.073	1.073
2-period change	0.813	0.646	0.613	0.604	1.361	1.345	1.338	1.338
3-period change	1.065	0.763	0.724	0.711	1.528	1.481	1.473	1.473
4-period change	1.279	0.830	0.800	0.784	1.645	1.569	1.563	1.562
5-period change	1.363	0.880	0.847	0.829	1.691	1.614	1.607	1.606
6-period change	1.390	0.910	0.873	0.855	1.708	1.637	1.628	1.628
7-period change	1.388	0.930	0.884	0.865	1.710	1.646	1.637	1.636
8-period change	1.353	0.932	0.884	0.860	1.701	1.645	1.635	1.634
9-period change	1.255	0.912	0.854	0.832	1.671	1.624	1.614	1.614
10-period change	1.154	0.895	0.824	0.806	1.641	1.606	1.595	1.595
11-period change	1.061	0.883	0.795	0.782	1.614	1.590	1.578	1.578
12-period change	0.992	0.883	0.767	0.761	1.593	1.577	1.563	1.563

6.3.2 Comparison of Alternative Estimators and Rotation Designs

The variances of alternative estimators under various rotation designs are given in Table 2. All variances are relative to the variance of the basic estimator of current level under that design. The efficiencies of alternative estimators of current level, change in level, and average level for multiple time periods are compared under the intermittent 4-8-4 rotation design and two continuous rotation designs. The continuous rotation designs are the 6-continuous scheme and the 8-continuous scheme. The 6-continuous scheme is the rotation scheme used in the Canadian Labor Force Survey conducted by Statistics Canada. For each period of the survey, the sample consists of six rotation groups, one rotation group in its first time-in-sample, ..., and one rotation group in its sixth time-in-sample. A given rotation group remains in the sample for six consecutive periods and then permanently drops out of the sample. See Kumar and Lee (1983) for more details about the design of the Canadian Labor Force Survey. In the 8-continuous scheme, there are 8 rotation groups in the sample for each period. A given rotation group remains in the sample for eight consecutive periods and then permanently drops out of the sample.

We compare the performance under the various rotation designs using the BLUE of current level based on 36 periods. We call this estimator the "best estimator" because its efficiency is vitually the same as that of the RRE. For all rotation schemes under consideration, there are some improvements in the precision of the estimators of current level from using the best estimator relative to the CPS composite estimator. As seen in Table 2, the gain is highest for employed where, under the 4-8-4 rotation scheme, the variance of the best estimator of current level is only 76% of that of the CPS composite estimator.

The precision of the estimators of change relative to the precision of the CPS composite estimator depends on the rotation design. From Table 2, we see that under the 4-8-4 rotation scheme, there is some gain in precision, which increases as the lag increases. For employed, the variance of the least squares estimator is 85% of the variance of the CPS composite estimator for one-period change, 61% of the variance of the CPS composite estimator for six-period change, and 76% of the variance of the CPS composite estimator for 12-period change. (Compare columns (2) and (3) of Table 2.)

Table 2
Variances of alternative estimators and rotation designs; the variance of the basic estimator of current level under each design equals one

	Employed				Unemployed			
Parameter	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.653	0.761	0.759	0.947	0.918	0.944	0.938
1-period change	0.511	0.432	0.395	0.434	1.070	1.073	1.003	1.051
2-period change	0.813	0.604	0.559	0.619	1.361	1.338	1.250	1.312
3-period change	1.065	0.710	0.669	0.747	1.528	1.473	1.372	1.443
4-period change	1.279	0.783	0.731	0.829	1.645	1.562	1.473	1.543
5-period change	1.363	0.828	0.782	0.901	1.691	1.606	1.533	1.607
6-period change	1.390	0.854	0.828	0.970	1.708	1.628	1.577	1.655
7-period change	1.388	0.863	0.874	1.026	1.710	1.636	1.612	1.686
8-period change	1.353	0.858	0.828	0.960	1.701	1.934	1.642	1.705
9-period change	1.255	0.830	0.960	1.108	1.671	1.614	1.663	1.719
10-period change	1.154	0.803	0.993	1.139	1.641	1.595	1.678	1.727
11-period change	1.061	0.779	1.021	1.165	1.614	1.578	1.688	1.733
12-period change	0.992	0.758	1.046	1.186	1.593	1.564	1.696	1.737
12-period average	0.369	0.326	0.440	0.394	0.255	0.249	0.301	0.266
12-change in averages	0.248	0.162	0.365	0.403	0.273	0.262	0.372	0.359

For estimating 12-period averages in employed using the 4-8-4 design, the CPS composite estimator is about 13% less efficient than the least squares estimator and, for estimating change in 12-period averages, it is about 53% less efficient, as can be seen by comparing the second and third columns of Table 2. For unemployed and the 4-8-4 design, there are only modest gains in precision from using the least squares estimator relative to the CPS composite estimator, as shown in the sixth and seventh columns of Table 2.

For estimation of 12-period change, 12-period average and change in 12-period averages, the 4-8-4 design is much superior to both continuous rotation designs for both characteristics. The continuous designs are generally superior for period-to-period changes for short periods.

6.3.3 Internal Consistency

In our analysis, we have constructed the best estimator of employed using only the past history of employed and the best estimator of unemployed using only the past history of unemployed. There is no formal reason not to include the past history of both employed and unemployed in the construction of the estimators. However, Fuller *et al.* (1993) state that the estimated cross correlations are less than 0.10, suggesting that there is little gain from such inclusion.

A method of constructing estimates of multiple characteristics that are internally consistent was suggested by Fuller (1990). In this procedure, estimates of employed, unemployed, and not-in-the-labor-force are constructed. Then these estimates are used as controls in a regression procedure to construct weights for the current observations. The weights can then be used to construct internally consistent estimates of any parameter of interest. The estimation procedure, including estimates of subdivisions of the labor force, is planned for implementation in 1998 for the CPS. See Lent, Miller and Cantwell (1996).

6.4 Conclusions

The main conclusions emerging form the variance computations in this section can be summarized as follows:

1. For all rotation designs and all characteristics under consideration, there are alternative estimation procedures with a variance of the current level smaller than that of the CPS composite estimator.
2. For estimation of change under the 4-8-4 rotation design, the gain in precision of the alternative estimators relative to the CPS composite estimator increases as the lag increases, and peaks around the lag of minimum overlap.
3. The intermittent 4-8-4 rotation design is inferior to the continuous rotation designs for short-period changes, but superior for current level, long-period averages, and changes in long-period averages.
4. The CPS composite estimator is comparable to the RRE for unemployed for the estimation of one-period change and 12-period change. However, the recursive regression estimation procedure is superior to the CPS composite estimator for other measures of change.
5. The RRE is more efficient in estimating change in level at lags for which the CPS composite estimator is not targeted, for instance, lags of four months to six months.

ACKNOWLEDGMENTS

The authors thank the referee and John Eltinge for helpful comments on earlier versions of this paper. Research for this paper was partly supported by Joint Statistical Agreement 91-21 with the U.S. Bureau of the Census and Cooperative Agreement 43-3AEU-80088 with the National Agricultural Statistics Service and the U.S. Bureau of the Census. The views expressed in this paper are those of the authors and do no necessarily represent the policies of the Bureau of the Census, the Bureau of Labor Statistics, or the National Agricultural Statistics Service.

APPENDIX

Lemma 1. Let the assumptions of the theorem hold. Then the variance of the estimator of current level θ_c converges to a positive number as the number of periods increases.

Proof. If the means $\theta_{c-1}, \theta_{c-2}, ..., \theta_{c-m}$, were known, then g_{1c}, $i = 1, 2, ..., s$ are unbiased estimators of θ_c, where $g_{1c} = y_{1,c}$; $g_{2,c} = y_{2,c} - b_{21}(y_{2,c-1} - \theta_{c-1})$; ...; and $g_{sc} = y_{sc} - \sum_{j=1}^{m} b_{sj}(y_{s,c-j} - \theta_{c-j})$. Furthermore, g_{1c}, $i = 1, 2, ..., s$ are independent with variances σ_i^2, $i = 1, 2, ..., s$. We may write the linear model:

$$g = J_s \theta_c + e \tag{A1}$$

where $g = (g_{1c}, g_{2c}, ..., g_{sc})'$, J_s is the $s \times 1$ column vector of ones, and e is the $s \times 1$ vector of errors with $E(e) = 0$, and $E(ee') = V_s = \text{Diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_s^2\}$. Thus the BLUE of θ_c for model (A1) has variance $(\sum_{i=1}^{s} \sigma_i^{-2})^{-1}$. By assumption, the variances σ_i^2, $i = 1, 2, ..., s$ are bounded below and the quantity $(\sum_{i=1}^{s} \sigma_i^{-2})^{-1}$ is a positive lower bound for the variance of the estimator of θ_c [see Lemma 4.2.3 of Yansaneh (1992)]. The variance of the estimator of θ_c is non-increasing as the number of observations increases, and hence, the variance converges to a positive number.

Lemma 2. Let the assumptions of the theorem hold. Then the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1}, ..., \theta_{t-1}$, based on data through time t, converges to a positive number as t increases.

Proof. First, suppose at a fixed time τ, at least m periods of observations are available both prior to τ and after τ. Define a transformation of the following form for the observations in each of the s streams at time τ: $u_{i\tau} = y_{i,\tau} - \sum_{j=-m}^{m} b_{k(i,\tau),j} y_{i,\tau-j}$, where $b_{k(i,\tau),0} = 0$ and $u_{i\tau}$ is uncorrelated with all observations preceding and succeeding $y_{i,\tau}$ in the i-th stream. Let the variance of $u_{i\tau}$ be λ_i^2, $i = 1, 2, ..., s$. These variances are bounded below by assumption. We conclude, as before, that there is a positive lower bound for the diagonal elements of the covariance matrix of the vector of recursive least squares estimators.

Now, assume that at time t, we begin the sequence of estimation with the vector of recursive least squares estimators $\hat{\Theta}_{t-1(m)} = (\hat{\theta}_{t-m}, ..., \hat{\theta}_{t-1})'$ based on data for the preceding m periods; and the vector of transformed observations $z_t' = (z_{1t}, ..., z_{st})$. Thus the linear model for the data at time t is given by (7), with c replaced by t. The data vector Z_t is of fixed dimension. Therefore, the covariance matrix of the BLUE of the vector of parameters $\Theta_{t(m+1)} = (\theta_{t-m}, ..., \theta_{t-1}, \theta_t)'$ is $\sum_{t(m+1)} = (W' V_t^{-1} W)^{-1}$. For computational convenience, we express W as $(I_{m+1}, M')'$, where I_{m+1} is the identity matrix of order $m+1$, and M is an $(s-1) \times (m+1)$ matrix which is constant over time. Thus we have

$$\begin{aligned} \textstyle\sum_{t(m+1)} &= (\Omega_{t-1(m+1)}^{-1} + M' \Omega_{00}^{-1} M)^{-1} \\ &= \Omega_{t-1(m+1)} - \Omega_{t-1(m+1)} M' D_t^{-1} M \Omega_{t-1(m+1)} \end{aligned} \tag{A2}$$

where

$\Omega_{t-1(m+1)} = \text{blockdiag}\{\sum_{t-1(m)}, \sigma_1^2\}$, $\Omega_{00} = \text{diag}\{\sigma_2^2, ..., \sigma_s^2\}$, and $D_t = \Omega_{00} + M \Omega_{t-1(m+1)} M'$. Since the second term on the right hand side of (A2) is positive definite, we conclude that the first m diagonal elements of $\sum_{t(m+1)}$ are less than or equal to the original diagonal elements of $\sum_{t-1(m)}$. This means that as t increases, the variances of the estimators of $\theta_{t-m}, ..., \theta_{t-2}, \theta_{t-1}$ are non-increasing. Since these variances are bounded below by a positive quantity, we conclude that the variances of the estimators of $\theta_{t-m}, ..., \theta_{t-2}, \theta_{t-1}$ converge to positive numbers as t increases.

Lemma 3. Let the assumptions of the theorem hold. Then, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, based on data through time t, converges to a positive number as t increases.

Proof. First, we show that variance of the least squares estimator of $\theta_c - \theta_{c-1}$ (where denotes the current period) converges as the number of periods increases by mimicking the arguments in the proof of Lemma 1. Also, arguments similar to those in the proof of Lemma 2 can be used to show that the variances of the least squares estimators of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, all converge as the number of periods increases.

Proof of theorem. Since $\sum_{t(m)}$ is a submatrix of the covariance matrix $\sum_{t(m+1)}$ of the least squares estimators of the full set of parameters $\theta_{t-m}, \theta_{t-m+1}, ..., \theta_{t-1}, \theta_t$, at time t, it is enough to show that $\sum_{t(m+1)}$ converges to a positive definite matrix as $t \to \infty$. From Lemma 1 and Lemma 2, each of the diagonal elements of $\sum_{t(m+1)}$ converges to a positive number as $t \to \infty$. From Lemma 3, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, ..., \theta_t - \theta_{t-1}$, converges to a positive number as $t \to \infty$. It follows that for each j, $1 \leq j \leq m$, the covariance between the least squares estimators of θ_t and θ_{t-j} converges as $t \to \infty$ and hence the covariance matrix $\sum_{t(m+1)}$ converges as $t \to \infty$.

Next, we prove that the limiting covariance matrix is positive definite. Let $\lim_{t\to\infty} \sum_{t(m)} = \sum_{(m)}$. It is enough to show that the variance of any non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 1, 2, ..., m$, is bounded below by a positive quantity. Let v_{mm} be the lower bound of every linear combination of the observations with one of the coefficients equal to one. The bound is positive by the assumption that the elements of V_n^{-1} are bounded.

Now, every estimator of the parameter θ_{t-j}, $j = 0, 1, ..., m$ is a linear combination of all observations such that the sum of the coefficients for the observations in the s streams at time $t-j$ is one, and the sum of the coefficients for the observations in the s streams at any other time is zero. This is a condition for the unbiasedness of the estimator for time $t-j$. For the sum of the coefficients of the s observations at time $t-j$ to be equal to one, at least one of the coefficients must be greater than or equal to s^{-1}. The minimum variance of any linear combination with first coefficient equal to s^{-1} is $s^{-2}v_{mm}$. Therefore, for $j = 0, 1, ..., m$, $\mathrm{Var}\{\hat{\theta}_{t-j}(t)\} \geq s^{-2}v_{mm}$.

Now, consider an arbitrary, non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 0, 1, ..., m$, given by $\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t)$, where, without loss of generality, $\gamma_0 = 1$. This linear combination can be expressed as

$$\begin{aligned}\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t) &= \hat{\theta}_t(t) + \sum_{j=1}^{m} \gamma_j \hat{\theta}_{t-j}(t) \\ &= \sum_{i=1}^{s} \sum_{h=1}^{t} c_{ih} y_{i,h} + \sum_{j=1}^{m} \gamma_j \sum_{i=1}^{s} \sum_{h=1}^{t} f_{ih(t-j)} y_{i,h} \\ &= \sum_{i=1}^{s} \left[c_{it} + \sum_{j=1}^{m} \gamma_j f_{it(t-j)}\right] y_{i,t} + \sum_{i=1}^{s} \sum_{h=1}^{t-1} \left[c_{ih} + \sum_{j=1}^{m} \gamma_j f_{ih(t-j)}\right] y_{i,h}\end{aligned} \tag{A3}$$

where c_{it}, $i = 1, 2 ..., s$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_t(t)$, and $f_{it(t-j)}$, $j = 1, ..., m$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_{t-j}(t)$, $j = 1, ..., m$, respectively. Therefore, $\sum_{i=1}^{s} c_{it} = 1$, and $\sum_{i=1}^{s} f_{it(t-j)} = 0$, for $j = 1, ..., m$. Thus $\sum_{i=1}^{s} [c_{it} + \sum_{j=1}^{m} \gamma_j f_{it(t-j)}] = 1$. That is, in the linear combination (A3), the sum of the coefficients for the observations $y_{i,t}$, $i = 1, 2, ..., s$, at time t is one. Therefore, at least one of the coefficients is greater than or equal to s^{-1}. Hence, $\mathrm{Var}\{\sum_{j=0}^{m} \gamma_j \hat{\theta}_{t-j}(t)\} \geq s^{-2}v_{mm}$, and we conclude that $\sum_{(m)}$ is positive definite.

REFERENCES

ADAM, A., and FULLER, W.A. (1992). Covariance estimators for the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 586-591.

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

BELL, W.R., and HILLMER, S.C. (1990). The time series approach to estimation for periodic surveys. *Survey Methodology*, 16, 195-215.

BINDER, D.A., and DICK, J.P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

BREAU, P., and ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 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.

FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1993). Estimators for longitudinal surveys. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, Ottawa, Canada, 309-324.

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. *Journal of the Royal Statistical Society*, Series B, 42, 221-226.

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour 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. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 130-139.

ODELL, P.L., and LEWIS, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B 12, 241-255.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.

RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

SCOTT, A.J., SMITH, T.M.F, and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

SINGH, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

YANSANEH, I.S. (1992). Least Squares Estimation for Repeated Surveys. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, Iowa.

YANSANEH, I.S. (1997). Recursive regression estimation in the presence of time-in-sample effects. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 164-169.