BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION

SRD Research Report Number: Census/SRD/89/04

VARIANCE FORMULA FOR THE GENERALIZED COMPOSITE ESTIMATOR UNDER A LONGITUDINAL MULTI-LEVEL ROTATION PLAN

by

Patrick J. Cantwell
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Recommended: Lawrence Ernst

Report Completed: June 1989

Report Issued: June 2, 1989

Variance Formula for the Generalized Composite Estimator Under a Longitudinal Multi-Level Rotation PLan

ABSTRACT

In a previous research report (Cantwell 1988), we obtained variance formulae for the generalized composite estimator applied to surveys using a balanced one-level rotation plan. The Current Population Survey satisfies the requirements of that design. The Survey of Income and Program Participation and the National Crime Survey, however, operate under "multi-level" designs. In each month one of p different groups is interviewed. Respondents then answer questions referring to the previous p months. Currently neither the SIPP nor NCS uses composite estimation to measure characteristics of interest. We derive simple expressions for the variance of the generalized composite estimator of level, change in level, and average level over time under longitudinal multi-level designs. The results apply to a wide range of survey designs.

1. INTRODUCTION

In SRD Report No. 88-26 (Cantwell 1988), variance formulae are derived for the generalized composite estimator under rotation plans satisfying the following conditions. First the period of reference for any interview is the current time period only. Such a design has been labeled "one-level." In addition, the rotation plan must be balanced--in any period, a new rotation group enters the sample, and follows the same pattern of periods in and out of sample as every preceding group, before finally retiring from the sample. Rotation groups are allowed to leave the sample temporarily before returning. Examples of surveys employing balanced one-level rotation plans are the Current Population Survey of the U.S. Bureau of the Census and the Labour Force Survey of Statistics Canada.

Many surveys operate under different designs. In the Survey of Income and Program Participation (SIPP), one of four different groups is interviewed each month, and respondents supply information on the previous four months. Under a different design, the National Crime Survey (NCS) alternately interviews one of six panels in any month, each group reporting crimes which occurred in the prior six months. See Nelson, McMillen, and Kasprzyk (1984) and Bureau of Justice Statistics, NCS Report NCJ-111456 (1988) for further details on the designs of the SIPP and NCS, respectively. Adhering to terminology found in Wolter (1979), we call these "multi-level" rotation plans to indicate that several time periods before the current one are being referenced.

Variance formulae which apply to balanced one-level designs no longer hold for multi-level plans. Further considerations arise. Suppose that $x_{h,i}$ is an estimator of some characteristic for month h obtained from the ith group in sample. Does the variance of $x_{h,i}$ depend on how many months elapse before the group is interviewed? We think this is possible. Is the survey "longitudinal," that is, are groups in sample interviewed many times over an extended period? Or is it "rotationally balanced," in that groups are methodically entered, interviewed, and retired in a balanced pattern?

This paper extends the results of SRD Report No. 88-26 to multi-level rotation plans. Restricting our attention to certain types of designs, we derive the variances of generalized composite estimators of level and change in level. The formulae are determined for single time periods (such as months) and combinations (such as quarters or years).

To facilitate reading, the sequence and notation of this report follow that of SRD Report No. 88-26 wherever possible. In Section 2, we discuss multi-level rotation plans, and specify which types are the focus of this report. Notation and previously used definitions are reintroduced. Our main results are stated in Theorems 1 through 3.

These theorems are proved in Section 3. Section 4 carries a discussion of several topics of interest, from the usefulness of composite estimation in certain surveys, to the complexities of some designs not covered in this report. Finally, we provide in an appendix several diagrams which illustrate the process for an example where each group is interviewed every fourth month. Included are depictions of the rotation plan with group estimates for different periods of time, and some of the general mathematical structures required to fill in the formulae given in the theorems.

2. RESTRICTIONS, NOTATION AND RESULTS

In SRD Report No. 88-26, the term "rotation group" is used to denote the set of respondents who enter the sample in a particular time period. To avoid confusion, in multi-level designs we call the entire set of people who are interviewed in a given period a "panel." This terminology is consistent with NCS, which employs six panels. On the other hand, the SIPP uses the label "rotation group" here, and calls the collection of these groups a "panel." We avoid the term rotation group, for reasons to be seen shortly, and use the word panel as defined above.

Throughout this report, we will use "month" to denote the period of time (i) in which interviews are done, and (ii) about which information is obtained. This is the period used in SIPP and NCS. However, our theorems and results extend to any period of time. When data are compiled and/or released to the public, "months" are often combined into quarters (of the year) or years.

When considering one-level rotation plans, we allowed a rotation group to assume any sequence of inclusions and exclusions from the sample, provided the design was balanced. For a multi-level plan, however, because of recall bias, it makes little sense to allow "design gaps," intermediate periods which are never referenced.

Consider an NCS panel which is interviewed in May and November. In November each respondent is asked about events or situations in May, June, July, August, September and October. Confusion may arise over which events occurred in April, and which in May. However, the previous interview in May, referencing November through April, can help place these events in the proper month. NCS goes so far as to conduct a preliminary "bounding interview" for those entering the sample. The responses from this initial meeting are not included in NCS estimates, but help to eliminate events which occurred before the reference period of the survey.

Suppose instead that a panel is interviewed every eight months and asked about the previous six, leaving gaps of two months after each interview. If a respondent confuses events which occurred six or seven months ago, the interviewer has no record to help determine the proper month. For this reason, and because we are not acquainted with any multi-level surveys which incorporate design gaps, we will restrict our efforts to multi-level rotation plans where (i) the sample is made of p panels, (ii) each panel is interviewed every pth "month," and (iii) the period of reference is the previous p months.

At this point, the question of sample replacement must be addressed. Is each panel interviewed many times, with little or no concern for balancing the time-in-sample in an individual month? In any SIPP sample, each of the four panels (that is, SIPP "rotation groups") is interviewed every fourth month through eight interviews, a period of almost three years. We might call such a design longitudinal, in that the panels could remain in sample indefinitely, and no attempt is made to balance any month's time-in-sample.

The design used in NCS, on the other hand, might be labeled "rotationally balanced." Each of the six panels is interviewed seven times, including the bounding interview. Within any panel there are seven rotation groups (although the group in sample for the first time is not used in the estimation process), making a total of 42 panel-rotation groups in sample at any time. After each interview, the rotation group which has just been interviewed for the seventh time leaves the sample, and a new one enters, so that data from any interview is balanced with respect to time-in-sample. The Consumer Expenditure Quarterly Survey uses a similar balanced design--each of three panels consists of five rotation groups (one is in sample only for "bounding" purposes).

Rotationally balanced multi-level designs are obviously more involved. For any month estimates are available (eventually) from each of the rotation groups in each of the panels. Realistic assumptions regarding the covariance structure and the various ways of combining these estimates grow more complex, and will be addressed in a later report. Here we consider only "longitudinal" designs with p panels. For any month a single estimate is eventually obtained from each panel. Effects of time-in-sample, including bias, will not be considered. This is not to imply that a rotationally balanced design will not supply longitudinal information, only that the model we consider here is simpler.

The interview of a panel will refer to the collective gathering of information in the assigned month from all sample units in that panel. For a particular characteristic which is to be estimated, let $x_{h,i}$ denote the estimate of "monthly" level <u>for month h</u> from the panel which is interviewed in month $h+i$, where $i = 1, 2, \ldots, p$. It is clear that i measures the recall time, that is, the amount of time between the interview and the month of reference. In the appendix is a chart depicting the estimates $x_{h,i}$ for a four-panel four-level design. In the diagram solid lines separate estimates which are obtained in different interviews. The SIPP refers to these boundaries between the reference periods of consecutive interviews as "seams."

Using this notation, $x_{h,1}, x_{h,2}, \ldots, x_{h,p}$ represent p estimates for month h obtained from the p panels in different interviews. On the other hand, $x_{h,p}, x_{h+1,p-1}, \ldots, x_{h+p-1,1}$ denote estimates for p different months obtained from one panel in a single interview.

The generalized composite estimator (GCE) for "monthly" level is defined recursively as follows:

$$y_h = \sum_{i=1}^{p} a_i x_{h,i} - k \sum_{i=1}^{p} b_i x_{h-1,i} + k y_{h-1} , \tag{1}$$

where k, the a_i's and the b_i's may take any values subject to $0 \leq k < 1$, $\sum_{i=1}^{p} a_i = 1$, and $\sum_{i=1}^{p} b_i = 1$. At this time, neither the SIPP nor NCS employs composite estimation. Each uses a simple average of the estimators (with appropriate adjustments) from the several panels for any given period of time. For greater detail on the GCE and how it compares in definition and computationally to other linear estimators, see Breau and Ernst (1983).

As in the case of a one-level design, the covariance structure of the monthly panel estimators here is assumed to be stationary in time. But now the effect of recall time on response enters. It may be reasonable to assume that response variability changes, in fact, likely increases, with the amount of time between the interview and the point of reference. We postulate the following covariance structure:

(i) $\text{Var}(x_{h,i}) = d_i^2 \sigma^2$ for all h and i, where $d_i > 0$;

(ii) $\text{Cov}(x_{h,i}, x_{h,j}) = 0$ for $i \neq j$, i.e., estimates for the same month from different panels are uncorrelated; and

(iii) For $r \geq 0$: $\text{Cov}(x_{h,i}, x_{h-r,j}) = \rho_{r,i} d_i d_j \sigma^2$, if the two x's refer to the same panel r months apart; or 0, otherwise. Take $\rho_{0,i}$ to be 1 for all i. (2)

It may well be that $d_1 \leq d_2 \leq \ldots \leq d_p$, if response variability increases with recall time. For the correlation coefficient $\rho_{r,i}$, r counts the number of months between estimates $x_{h,i}$ and $x_{h-r,j}$. The index i indicates that the estimate for month h is recorded from an interview in month $h+i$. It may appear as if the subscript j in $x_{h-r,j}$ plays no part in determining $\mathrm{Cov}(x_{h,i}, x_{h-r,j})$. However, there is only one value j, $1 \le j \le p$, for which the estimates $x_{h,i}$ and $x_{h-r,j}$ refer to the same panel. (This value is $j = \mathrm{mod}_p(i+r-1) + 1$, where $\mathrm{mod}_p(n)$ is the value of the integer n, modulo p.) Otherwise, the covariance is 0.

The coefficients $\rho_{r,i}$ will likely decrease in r for fixed i, reflecting smaller correlation as the separation between points in time grows. The effect of varying i for fixed r, though, is harder to predict, and may be related to the survey conducted and the characteristic being enumerated. In some cases, it may be appropriate to replace $\rho_{r,1}, \rho_{r,2}, \ldots, \rho_{r,p}$ with a commom ρ_r. Alternatively, the values of the $\rho_{r,i}$'s for different i's could depend on how many times the relevant panel has been interviewed between months $h-r$ and h. Results will be stated with general correlation coefficients $\rho_{r,i}$; the reader can make substitutions according to his model or experience.

Define the vectors a and b as $(a_1, a_2, \ldots, a_p)^T$ and $(b_1, b_2, \ldots, b_p)^T$, respectively, from the coefficients in the GCE. The symbol I denotes the $p \times p$ identity matrix. Let D be the $p \times p$ diagonal matrix with $d_1, d_2, \ldots, d_p$ down the diagonal. Similarly, for any $r \ge 0$, let R_r be the $p \times p$ diagonal matrix with $\rho_{r,1}, \rho_{r,2}, \ldots, \rho_{r,p}$ down the diagonal. Define the $p \times p$ matrix J by: $J_{i,i+1} = 1$ for $i = 1, 2, \ldots, p-1$; $J_{p1} = 1$; and $J_{ij} = 0$, otherwise. The general forms of the matrices J, D, and R_r are shown in the appendix for $p = 4$. Finally, let

$$Q = \sum_{n=1}^{\infty} k^n R_n J^n. \tag{3}$$

In Section 3 we will prove that the sum in (3) converges.

We state several theorems, and leave the proofs to Section 3. All results apply to the limiting case where panels have been in sample long enough to eliminate the effect of phasing in the sample. If the $\rho_{r,i}$'s decrease rapidly with r, or if k is relatively small, the "steady-state" arrives soon. This point is discussed in greater detail in Section 4.

Theorem 1. If the GCE of level is defined as in (1), and the covariance structure of (2) holds, then

$$\mathrm{Var}(y_h) = \sigma^2 \{a^T D^2 a + k^2 b^T D^2(b-2a) + 2(a-k^2 b)^T DQD(a-b)\} / (1-k^2) \tag{4}$$

Notice that when one uses an unweighted average of the estimates for month h from the p panels, $k = 0$, $Q = 0$, and $a_i = 1/p$ for $i = 1, 2, \ldots, p$. Then $\mathrm{Var}(y_h) = (\sigma^2/p^2) \sum_{i=1}^{p} d_i^2$, as expected.

Theorem 2. Let $y_h - y_{h-1}$ be the GCE estimator of "monthly" change.

(i) If $k = 0$, then $\mathrm{Var}(y_h - y_{h-1}) = 2\sigma^2 a^T D(I-R_1 J)Da$;

(ii) if $0 < k < 1$, then $\mathrm{Var}(y_h - y_{h-1})$

$$= \sigma^2\{a^T D^2 a + k^2 b^T D^2 b - 2ka^T DR_1 JDb\}/k - (1-k)^2 \mathrm{Var}(y_h)/k \tag{5}$$

Often of interest are the average over a certain period of time, for example, a quarter or a year, the difference in these averages from one period to the next, or even the difference in "monthly" level for two months a year apart. Denote by $S_{h,t}$ the sum of the GCE's <u>for the last t months</u>: $S_{h,t} = y_h + y_{h-1} + \cdots + y_{h-t+1}$, $t \geq 1$. Note that $S_{h,t}$ is defined slightly differently here than in SRD Report No. 88-26. Commonly used values of t include three, four, and twelve. We will leave it to the user to divide $S_{h,t}$ by t if he desires an average rather than a sum.

Theorem 3. The expressions $S_{h,t}$, $y_h - y_{h-t}$, and $S_{h,t} - S_{h-t,t}$ can be written as $\sum_{i=0}^{\infty} v_i^T x_{h-i}$, where, for any of these expressions, v_0, v_1, v_2, ... is a sequence of $p \times 1$ nonrandom vectors. In particular,

(i) for $S_{h,t}$:

$$v_i = \begin{cases} a + [(k - k^{i+1})/(1-k)](a-b), & i = 0, 1, \ldots, t-1, \\ [k^{i-t}(k - k^{t+1})/(1-k)](a-b), & i = t, t+1, t+2, \ldots; \end{cases}$$

(ii) for $y_h - y_{h-t}$: $v_0 = a$, $v_t = k^t(a-b) - a$, and

$$v_i = \begin{cases} k^i(a-b), & i = 1, 2, \ldots, t-1, \\ -k^{i-t}(1 - k^t)(a-b), & i = t+1, t+2, t+3, \ldots; \text{ and} \end{cases}$$

(iii) for $S_{h,t} - S_{h-t,t}$:

$$v_i = \begin{cases} a + [(k - k^{i+1})/(1-k)](a-b), & i = 0, 1, \ldots, t-1, \\ [(2k^{i-t+1} - k - k^{i+1})/(1-k)](a-b) - a, & i = t, t+1, \ldots, 2t-1, \\ -[k^{i-2t+1}(1-k^t)^2/(1-k)](a-b), & i = 2t, 2t+1, \ldots \end{cases}$$

In each of the three cases, $\mathrm{Var}(\sum_{i=0}^{\infty} v_i^{\mathrm{T}} x_{h-i})$

$$= \sigma^2 \{ \sum_{i=0}^{\infty} v_i^{\mathrm{T}} D^2 v_i + 2 \sum_{i=0}^{\infty} v_i^{\mathrm{T}} \sum_{n=1}^{\infty} DR_n J^n D\, v_{i+n} \} \tag{6}$$

The sums in (6) converge because, in the three parts of the theorem, v_i is proportional to $k^i(a-b)$ for $i \geq 2t$.

3. DERIVATIONS OF THE THEOREMS

The proofs of the theorems just stated sometimes resemble the proofs of related theorems in SRD Report No. 88-26. We have retained the similarities wherever possible to emphasize how closely the two designs are structured, and to facilitate the reading of the proofs by one who is familiar with the former report.

The letter p is used to denote the number of panels as well as the number of months about which information is obtained during any interview. J is the $p \times p$ matrix with 1's for J_{12}, J_{23}, $\ldots$, $J_{p-1,p}$, and $J_{p,1}$, and 0's everywhere else. It is easily seen that, if U is any $n \times p$ matrix comprising p $n\times 1$ vectors $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p)$, then $\mathbf{U}\mathbf{J} = (\mathbf{u}_p, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{p-1})$. That is, postmultiplication by $\mathbf{J}$ moves the last column to the front, and moves each remaining column back one position. Therefore, the product $\mathbf{J}^2$ has 1's for components (1,3), (2,4), ..., $(p-2,p)$, $(p-1,1)$, and $(p,2)$. The pattern for general r follows. The pth power of $\mathbf{J}$ is $\mathbf{J}^p = \mathbf{I}$, and the cycle begins again with $\mathbf{J}^{p+1} = \mathbf{J}$. The form of $\mathbf{J}$ and $\mathbf{J}^r$ are illustrated in the appendix for $p = 4$.

Vectors are formed out of the estimates obtained from the different panels referring to the same month. For any month h, let $\mathbf{x}_h$ be the $p\times 1$ vector $(x_{h,1}, x_{h,2}, \ldots, x_{h,p})^{\mathrm{T}}$ of estimates. $\mathbf{D}$ was defined as the $p\times p$ matrix with $d_1, d_2, \ldots, d_p$ down the diagonal and 0's elsewhere. The first two parts of (2) giving the covariance structure of the estimates imply that $\mathrm{Var}(\mathbf{x}_h) = \sigma^2\mathbf{D}^2$ for all h.

For any $r \geq 0$, the $p\times p$ matrix $\mathbf{R}_r$ has correlation coeeficients $\rho_{r,1}, \rho_{r,2}, \ldots, \rho_{r,p}$ on the diagonal, and 0's elsewhere. From part three of (2) we deduce that $\mathrm{Cov}(\mathbf{x}_h,\mathbf{x}_{h-1}) = \sigma^2\mathbf{D}\mathbf{R}_1\mathbf{J}\mathbf{D}$. This follows as the nonzero components of $\mathbf{J}$ indicate the pairs of estimates, one from month h and one from h-1, which arise from the same panel. The other matrices in the product ensure that (a) for $i = 1, 2, \ldots, p-1$, the $(i,i+1)$ component is $\sigma^2\rho_{1,i}d_id_{i+1}$, and (b) the $(p,1)$ component is $\sigma^2\rho_{1,p}d_pd_1$.

The matrix $\mathbf{J}^2$ was shown to have 1's for the components (1,3), (2,4), ..., $(p-2,p)$, $(p-1,1)$, and $(p,2)$. The same development as above implies $\mathrm{Cov}(\mathbf{x}_h,\mathbf{x}_{h-2}) = \sigma^2\mathbf{D}\mathbf{R}_2\mathbf{J}^2\mathbf{D}$. In general,

$$\mathrm{Cov}(\mathbf{x}_h,\mathbf{x}_{h-r}) = \sigma^2\mathbf{D}\mathbf{R}_r\mathbf{J}^r\mathbf{D}, \quad \text{for } r = 0, 1, 2, \ldots, \text{ and all } h. \qquad (7)$$

As r increases, $\mathbf{J}^r$ was seen to run through a cycle of p matrices, that is, it never dies out. However, the correlation coefficients in $\mathbf{R}_r$ may well approach 0 as r increases, and the covariance terms would become arbitrarily small.

In (3) the matrix Q was defined as $\sum_{n=1}^{\infty} k^n R_n J^n$. The ijth cell, Q_{ij}, is then an infinite sum of terms. Its pattern can be demonstrated by showing a couple of cells:

$$Q_{12} = k\rho_{1,1} + k^{p+1}\rho_{p+1,1} + k^{2p+1}\rho_{2p+1,1} + \dots = \sum_{m=0}^{\infty} k^{pm+1}\rho_{pm+1,1};$$

$$Q_{ii} = k^p\rho_{p,i} + k^{2p}\rho_{2p,i} + k^{3p}\rho_{3p,i} + \dots = \sum_{m=0}^{\infty} k^{pm+p}\rho_{pm+p,i}.$$

For general p, i, and j, we can write

$$Q_{ij} = \sum_{m=0}^{\infty} k^u\rho_{u,i}, \quad \text{where} \quad u = pm+1+\text{mod}_p(p-i+j-1).$$

To show that the sum in (3) converges, it suffices to show that the sum of terms in any cell Q_{ij} converges absolutely. Because the correlation coefficients $\rho_{r,i}$ are less than or equal to 1 in absolute value for any r and i, it is easily seen from the expression of Q_{ij} above that

$$|Q_{ij}| \leq \sum_{m=0}^{\infty} |k^u\rho_{u,i}| < \sum_{n=1}^{\infty} k^n,$$ which converges. If k or the correlations are small, the convergence will be rapid.

The generalized composite estimator was written in (1) as

$$y_h = \sum_{i=1}^{p} a_i x_{h,i} - k\sum_{i=1}^{p} b_i x_{h-1,i} + ky_{h-1}$$

Writing this in vector form and substituting repeatedly:

$$\begin{aligned} y_h &= a^T x_h - kb^T x_{h-1} + ky_{h-1} \\ &= a^T x_h - kb^T x_{h-1} + k(a^T x_{h-1} - kb^T x_{h-2} + ky_{h-2}) \\ &= a^T x_h + k(a-b)^T x_{h-1} - k^2 b^T x_{h-2} + k^2 y_{h-2} \\ &= a^T x_h + k(a-b)^T x_{h-1} - k^2 b^T x_{h-2} \\ &\quad + k^2(a^T x_{h-2} - kb^T x_{h-3} + ky_{h-3}) \end{aligned}$$

$$= a^T x_h + k(a-b)^T x_{h-1} + k^2(a-b)^T x_{h-2} - k^3 b^T x_{h-3} + k^3 y_{h-3}$$

$$= \dots$$

$$= a^T x_h + k(a-b)^T x_{h-1} + k^2(a-b)^T x_{h-2} + k^3(a-b)^T x_{h-3} + \cdots$$

$$= a^T x_h + (a-b)^T \sum_{i=1}^{\infty} k^i x_{h-i} \tag{8}$$

Proofs of the theorems follow from the results in (7) and (8).

PROOF OF THEOREM 1.

$$\begin{aligned}
\mathrm{Var}(y_h) &= a^T \mathrm{Var}(x_h) a + (a-b)^T \sum_{i=1}^{\infty} k^{2i} \mathrm{Var}(x_{h-i})(a-b) \\
&\quad + 2\, a^T \sum_{i=1}^{\infty} k^i \mathrm{Cov}(x_h, x_{h-i})(a-b) \\
&\quad + 2\, (a-b)^T \sum_{1 \le i < j}^{\infty} \sum^{\infty} k^{i+j} \mathrm{Cov}(x_{h-i}, x_{h-j})(a-b) \\
&= a^T \sigma^2 D^2\, a + (a-b)^T \sum_{i=1}^{\infty} k^{2i}\, \sigma^2 D^2\, (a-b) \\
&\quad + 2\, a^T \sum_{i=1}^{\infty} k^i\, \sigma^2 D R_i J^i D\, (a-b) \\
&\quad + 2\, (a-b)^T \sum_{i=1}^{\infty} k^{2i} \sum_{j=i+1}^{\infty} k^{j-i}\, \sigma^2 D R_{j-i} J^{j-i} D\, (a-b) \\
&= \sigma^2 \{\, a^T D^2 a + (a-b)^T D^2 (a-b)\, k^2/(1-k^2) \\
&\quad + 2\, a^T D\, [\sum_{i=1}^{\infty} k^i R_i J^i]\, D(a-b) \\
&\quad + 2\, (a-b)^T D \sum_{i=1}^{\infty} k^{2i}\, [\sum_{n=1}^{\infty} k^n R_n J^n]\, D(a-b) \,\}
\end{aligned} \tag{9}$$

Recall again that $Q = \sum_{n=1}^{\infty} k^n R_n J^n$. Both expressions in brackets in (9) are equal to Q. Using the fact that D is symmetric, line (9) can be rewritten as:

$$\sigma^2 \{\, a^T D^2 a + (a-b)^T D^2 (a-b)\, k^2/(1-k^2) + 2\, a^T D Q D(a-b) + 2\, (a-b)^T D Q D(a-b)\, k^2/(1-k^2) \,\}$$

$$= \sigma^2 \{ (1-k^2)a^T D^2 a + k^2(a^T D^2 a - 2b^T D^2 a + b^T D^2 b)$$
$$+ 2(1-k^2)\ a^T DQD(a-b)$$
$$+ 2k^2\ [a^T DQD(a-b) - b^T DQD(a-b)] \} / (1-k^2)$$
$$= \sigma^2 \{ a^T D^2 a + k^2 b^T D^2 (b-2a) + 2(a-k^2 b)^T DQD(a-b) \} / (1-k^2).$$

and the theorem is proved.

Proof of Theorem 2. If $k = 0$, $y_h = a^T x_h$ and $y_{h-1} = a^T x_{h-1}$. From prior results regarding the stationarity of the covariance,

$$\text{Var}(y_h) = \text{Var}(y_{h-1}) = a^T \sigma^2 D^2 a = \sigma^2 a^T D^2 a.$$

The estimator of "month-to-month" change is $y_h - y_{h-1} = a^T x_h - a^T x_{h-1}$. Its variance is

$$\text{Var}(y_h - y_{h-1}) = 2\sigma^2 a^T D^2 a - 2a^T \sigma^2 DR_1 JDa = 2\sigma^2 a^T D(I-R_1 J)Da.$$

If $0 < k < 1$, $y_h = a^T x_h - kb^T x_{h-1} + ky_{h-1} \equiv V_h + ky_{h-1}$, (10)

where V_h is defined as $a^T x_h - kb^T x_{h-1}$.

$$\text{Var}(V_h) = a^T \sigma^2 D^2 a + k^2 b^T \sigma^2 D^2 b - 2ka^T \sigma^2 DR_1 JDb$$
$$= \sigma^2 \{ a^T D^2 a + k^2 b^T D^2 b - 2ka^T DR_1 JDb \} \qquad (11)$$

It follows from (10) that

$$\text{Var}(y_h) = \text{Var}(V_h) + k^2\text{Var}(y_{h-1}) + 2k\text{Cov}(V_h, y_{h-1})$$
$$= \text{Var}(V_h) + k^2\text{Var}(y_h) + 2k\text{Cov}(V_h, y_{h-1}),$$ and thus

$$2\text{Cov}(V_h, y_{h-1}) = (1/k) \{ (1-k^2)\text{Var}(y_h) - \text{Var}(V_h) \}$$

Now we can write $y_h - y_{h-1} = V_h + ky_{h-1} - y_{h-1} = V_h - (1-k)y_{h-1}$.

$$\text{Var}(y_h - y_{h-1}) = \text{Var}(V_h) + (1-k)^2\text{Var}(y_{h-1}) - 2(1-k)\text{Cov}(V_h, y_{h-1})$$
$$= \text{Var}(V_h) + (1-k)^2\text{Var}(y_h)$$
$$- (1-k)(1/k) \{ (1-k^2)\text{Var}(y_h) - \text{Var}(V_h) \}$$
$$= [1 + (1-k)/k]\ \text{Var}(V_h) + [(1-k)^2 - (1/k)(1-k)(1-k^2)]\ \text{Var}(y_h)$$
$$= (1/k)\ \text{Var}(V_h) + (1/k)(1-k)^2[k - (1+k)]\ \text{Var}(y_h)$$
$$= \text{Var}(V_h)/k - (1-k)^2\text{Var}(y_h)/k$$

Substituting from (11) finishes the proof:

$$= \sigma^2\{ a^T D^2 a + k^2 b^T D^2 b - 2k a^T D R_1 J D b\}/k \quad - \quad (1-k)^2 \mathrm{Var}(y_h)/k$$

Proof of Theorem 3. To start, the variance of the general summation $\sum_{i=0}^{\infty} v_i^T x_{h-i}$ is easily derived.

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=0}^{\infty} v_i^T x_{h-i}\right) &= \sum_{i=0}^{\infty} v_i^T \mathrm{Var}(x_{h-i}) v_i + 2 \sum_{0 \le i < j}^{\infty} \sum^{\infty} v_i^T \mathrm{Cov}(x_{h-i}, x_{h-j}) v_j \\
&= \sum_{i=0}^{\infty} v_i^T \sigma^2 D^2 v_i + 2 \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} v_i^T \sigma^2 D R_{j-i} J^{j-i} D v_j \\
&= \sigma^2 \left\{ \sum_{i=0}^{\infty} v_i^T D^2 v_i + 2 \sum_{i=0}^{\infty} v_i^T \sum_{n=1}^{\infty} D R_n J^n D v_{i+n} \right\}
\end{aligned}$$

In the remainder of the proof, we need only show that the sums and differences specified in the theorem can be expressed as $\sum_{i=0}^{\infty} v_i^T x_{h-i}$, with the sequence of v_i's as given in parts (i) through (iii).

(i) $S_{h,t}$ was defined as $y_h + y_{h-1} + \cdots + y_{h-t+1}$, for any $t \ge 1$. The vectors $v_0, v_1, v_2, \ldots$ can be determined by introducing y_{h-1} terms one at a time.

$$\begin{aligned}
y_h + y_{h-1} &= a^T x_h - k b^T x_{h-1} + k y_{h-1} + y_{h-1} \\
&= a^T x_h - k b^T x_{h-1} + (1+k)(a^T x_{h-1} - k b^T x_{h-2} + k y_{h-2}) \\
&= a^T (x_h + x_{h-1}) + k(a-b)^T x_{h-1} - (k+k^2) b^T x_{h-2} + (k+k^2) y_{h-2}
\end{aligned}$$

Continuing,

$$\begin{aligned}
y_h + y_{h-1} + y_{h-2} &= a^T (x_h + x_{h-1}) + k(a-b)^T x_{h-1} - (k+k^2) b^T x_{h-2} \\
&\quad + (1+k+k^2)(a^T x_{h-2} - k b^T x_{h-3} + k y_{h-3}) \\
&= a^T (x_h + x_{h-1} + x_{h-2}) + k(a-b)^T x_{h-1} + (k+k^2)(a-b)^T x_{h-2} \\
&\quad - (k+k^2+k^3) b^T x_{h-3} + (k+k^2+k^3) y_{h-3}
\end{aligned}$$

Including all t terms,

$$y_h + \dots + y_{h-t+1} = a^T(x_h + \dots + x_{h-t+1}) + k(a-b)^T x_{h-1}$$
$$+ (k+k^2)(a-b)^T x_{h-2} + \dots + (k+k^2+ \dots + k^{t-1})(a-b)^T x_{h-t+1}$$
$$- (k+k^2+ \dots + k^t) b^T x_{h-t} + (k+k^2+ \dots + k^t) y_{h-t} \qquad (12)$$

But according to (8),

$$y_{h-t} = a^T x_{h-t} + (a-b)^T \sum_{i=1}^{\infty} k^i x_{h-t-i}$$

The last two terms of (12) become

$$(k+k^2+ \dots + k^t)[\,(a-b)^T x_{h-t} + (a-b)^T \sum_{i=1}^{\infty} k^i x_{h-t-i}\,]$$
$$= (k+k^2+ \dots + k^t)\,(a-b)^T \sum_{j=t}^{\infty} k^{j-t} x_{h-j}$$

Finally we can write (12) as

$$y_h + \dots + y_{h-t+1} = a^T(x_h + \dots + x_{h-t+1}) + k(a-b)^T x_{h-1}$$
$$+ (k+k^2)(a-b)^T x_{h-2} + \dots + (k+k^2+ \dots + k^{t-1})(a-b)^T x_{h-t+1}$$
$$+ (k+k^2+ \dots + k^t)(a-b)^T \sum_{j=t}^{\infty} k^{j-t} x_{h-j} \qquad (13)$$

Now it is apparent that: $v_0 = a$; $v_i = a + (k+k^2+ \dots + k^i)(a-b)$, for $i = 1, 2, \dots, t-1$; and $v_i = (k+k^2+ \dots + k^t)k^{i-t}(a-b)$, for $i = t, t+1, \dots$ The series of v_i given in Theorem 3, part (i) are obtained by summing powers of k.

(ii) The difference $y_h - y_{h-t}$ can be replaced by the appropriate summations:

$$y_h - y_{h-t} = a^T x_h + (a-b)^T \sum_{i=1}^{\infty} k^i x_{h-i} - [a^T x_{h-t} + (a-b)^T \sum_{i=1}^{\infty} k^i x_{h-t-i}]$$
$$= a^T x_h + (a-b)^T \,[\, \sum_{i=1}^{t-1} k^i x_{h-i} + k^t x_{h-t} + \sum_{i=t+1}^{\infty} k^i x_{h-i}\,]$$
$$- a^T x_{h-t} - (a-b)^T \sum_{j=t+1}^{\infty} k^{j-t} x_{h-j}$$

$$= \mathbf{a}^{\mathrm{T}}\mathbf{x}_h + (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \sum_{i=1}^{t-1} k^i \mathbf{x}_{h-i} + [k^t(\mathbf{a}-\mathbf{b}) - \mathbf{a}]^{\mathrm{T}}\mathbf{x}_{h-t}$$

$$+ (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \sum_{i=t+1}^{\infty} (k^i - k^{i-t})\mathbf{x}_{h-i} \tag{14}$$

From (14) the vectors $\mathbf{v}_i$ can be determined.

(iii) To find the $\mathbf{v}_i$'s corresponding to $S_{h,t} - S_{h-t,t}$, one need only combine the appropriate vectors from each separate sum, as written in part (i). In $S_{h,t}$ the vectors are:

$$\mathbf{v}_i = \begin{cases} \mathbf{a} + [(k - k^{i+1})/(1-k)](\mathbf{a}-\mathbf{b}), & i = 0, 1, \ldots, t-1, \\ [k^{i-t}(k - k^{t+1})/(1-k)](\mathbf{a}-\mathbf{b}), & i = t, t+1, t+2, \ldots \end{cases} \tag{15}$$

Notice that $S_{h-t,t} = \sum_{i=0}^{\infty} \mathbf{v}_i \mathbf{x}_{h-t-i} = \sum_{j=t}^{\infty} \mathbf{v}_{j-t}\mathbf{x}_{h-j}$. The same expressions can be used for $S_{h-t,t}$, except that the vector indices start at t rather than 0, and are shifted: $i = j-t$. In $S_{h-t,t}$ the vectors are:

$$\mathbf{v}_j = \begin{cases} \mathbf{a} + [(k - k^{j-t+1})/(1-k)](\mathbf{a}-\mathbf{b}), & j = t, t+1, \ldots, 2t-1, \\ [k^{j-2t}(k - k^{t+1})/(1-k)](\mathbf{a}-\mathbf{b}), & j = 2t, 2t+1, 2t+2, \ldots \end{cases}$$

Now the solution for $S_{h,t} - S_{h-t,t}$ can be found by subtracting appropriate terms from each sum. For $i = 0, 1, \ldots, t-1$, only the expression from (15) is necessary. When $i = t, t+1, \ldots, 2t-1$,

$$\begin{aligned} \mathbf{v}_i &= \{ [k^{i-t}(k - k^{t+1}) - (k - k^{i-t+1})]/(1-k) \}(\mathbf{a}-\mathbf{b}) - \mathbf{a} \\ &= [(2k^{i-t+1} - k - k^{i+1})/(1-k)](\mathbf{a}-\mathbf{b}) - \mathbf{a} \end{aligned}$$

Finally, for $i = 2t, 2t+1, 2t+2, \ldots$,

$$\begin{aligned} \mathbf{v}_i &= \{ [k^{i-t}(k - k^{t+1}) - k^{i-2t}(k - k^{t+1})]/(1-k) \}(\mathbf{a}-\mathbf{b}) \\ &= -[k^{i-2t}(1 - k^t)k(1 - k^t)/(1-k)](\mathbf{a}-\mathbf{b}) \\ &= -[k^{i-2t+1}(1 - k^t)^2/(1-k)](\mathbf{a}-\mathbf{b}) \end{aligned}$$

4. ADDITIONAL COMMENTS

Several unrelated topics are discussed in this section. Of primary importance is how useful these results are in actual surveys. In SRD Report No. 88-26, the examples mentioned include the Current Population Survey and Statistics Canada's Labour Force Survey. Each survey gathers data on labor force characteristics, such as work force and employed status. The correlations between rotation group estimates from one month to the next tend to be moderately positive, and beneficial to the implementation of composite estimation. Of course, the developments in the report apply to any survey employing a balanced one-level rotation design.

The SIPP and NCS are the two examples mentioned most frequently in this report. Many of the characteristics measured in the NCS involving incidents of crimes may exhibit a negligible correlation in the panel estimates from one month to the next. If so, it would appear questionable whether the NCS could profit by using composite estimation rather than simple linear estimation from the months involved. We offer no argument here to the contrary. This point will be addressed in a subsequent report dealing with rotationally balanced multi-level rotation plans, which include the NCS design.

On the other hand, the SIPP seeks information on income level, sources of income, program participation, and other items. For many of these, the correlations of interest may be large enough to make our results useful to the SIPP. As always, the theorems have been put in a form to be applicable to general surveys.

A second matter to consider is the "steady-state" of affairs mentioned before the statements of the theorems in Section 2. By this we mean the limiting case where panels have been in sample long enough to eliminate the affect of phasing in the sample. Our claim is that assuming a "steady-state" from the beginning of the survey usually will not change the true variances by much.

Consider as an example a survey where each of four panels are interviewed every fourth month through eight interviews. The SIPP uses such a design. Data is accumulated from some or all of the panels for 35 months. Suppose k is assigned a value of 0.5, and a composite estimate is desired for month 7. The first panel (note: the SIPP would call this the first rotation group) contributes seven months of data up to this point: $x_{1,4}$, $x_{2,3}$, $x_{3,2}$, $x_{4,1}$, $x_{5,4}$, $x_{6,3}$, and $x_{7,2}$. The last three estimates are not available until the interview in month 9. (See the chart in the appendix.) Similarly, from the fourth panel, estimates $x_{4,4}$, $x_{5,3}$, $x_{6,2}$, and $x_{7,1}$ are obtained during an interview conducted in month 8.

The derivation in (8), applied to month 7, starts with

$y_7 = a^T x_7 - kb^T x_6 + ky_6$, and concludes with

$$y_7 = a^T x_7 + (a-b)^T \sum_{i=1}^{\infty} k^i x_{7-i}.$$

Obviously there are no vectors x_{7-i} for i greater than 6. In fact, only partial vectors are obtained for i = 4, 5, or 6. One remedy is to change the weights a and b, but only for x_1, x_2, and x_3.

Let $x_1 = (0, 0, 0, x_{1,4})^T$, $x_2 = (0, 0, x_{2,3}, x_{2,4})^T$, and $x_3 = (0, x_{3,2}, x_{3,3}, x_{3,4})^T$, the estimates available from months 1, 2, and 3. We define special coefficient vectors: $a_1 = (0, 0, 0, a_{14})^T$, $b_1 = (0, 0, 0, b_{14})^T$, $a_2 = (0, 0, a_{23}, a_{24})^T$, $b_2 = (0, 0, b_{23}, b_{24})^T$, $a_3 = (0, a_{32}, a_{33}, a_{34})^T$, $b_3 = (0, b_{32}, b_{33}, b_{34})^T$. In order to ensure that the estimators in all months are unbiased (if we ignore time-in-sample bias), we require that $\sum_{j=1}^{4} a_{ij} = \sum_{j=1}^{4} b_{ij} = 1$, for i = 1, 2, and 3. It follows immediately that $a_{14} = b_{14} = 1$. Beyond the third month, a and b are selected as usual. Then

$$y_7 = a^T x_7 + (a-b)^T \sum_{i=1}^{3} k^i x_{7-i} + k^4 (a_3 - b_3)^T x_3 + k^5 (a_2 - b_2)^T x_2 \tag{16}$$

The only adjustment necessary in the variance formulae is to amend variances and covariances corresponding to x_2, x_3, and "nonexistent" terms. In (9), the expression

$$a^T \mathrm{Var}(x_h) a + (a-b)^T \sum_{i=1}^{\infty} k^{2i} \mathrm{Var}(x_{h-i})(a-b)$$
$$= \sigma^2 \{ a^T D^2 a + (a-b)^T D^2 (a-b)\ k^2/(1-k^2) \}$$

is very close to the actual sum of variances components from (16):

$$\sigma^2 \{ a^T D^2 a + (k^2 + k^4 + k^6)(a-b)^T D^2 (a-b) + k^8 (a_3 - b_3)^T D^2 (a_3 - b_3)$$
$$+ k^{10} (a_2 - b_2)^T D^2 (a_2 - b_2) \}$$

Two expressions in (9) contain the matrix $Q = \sum_{n=1}^{\infty} k^n R_n J^n$, defined as if all x_{7-i} were sampled. It is not difficult to see that the actual sum of covariance components will again be slightly different from that obtained with an infinite sum approximation. However multiplication by the correlation coefficients in R_n, generally a bit smaller than one, will reduce the relative difference even more.

In all, the infinite sums used in the variance formula for y_7 provide a good approximation to the actual variance. If a smaller value of k is used or a larger value of h is desired then this difference narrows even further.

Another aspect to consider is the covariance structure laid out in (2). Our experience has led us to expect that, as recall time increases, so does response variability. This conclusion may be reasonable in demographic surveys, where respondents often supply information from memory.

Nevertheless it has been pointed out to us that a somewhat opposite effect may occur in some business surveys. It may be the case that, for a while, response variability actually decreases with time. In some circumstances survey data are derived from records which may not be complete or sufficiently accurate for several months. Minimum variance in the responses might then be exhibited several months prior to the time of the interview.

Whether or not the variance of a panel estimator is a monotonic function of the recall time, the results in Section 2 are valid under the model presented. No assumptions are made about the constants $d_1, d_2, \ldots, d_p$, except that they are positive. Obtaining good approximations for the d_i's is the responsibility of the individual using these results.

A final point to raise is the difficulty of finding easily applied general formulae for a rotationally balanced multi-level rotation plan. Such a design is more symmetric than the longitudinal plan considered in this report in some aspects, including time-in-sample. For any month, estimates are eventually obtained from each panel, one panel recalling one month, another recalling two months, etc. Each panel comprises a set of rotation groups representing the entire range of times-in-sample.

This symmetry is offset, at least computationally, by the more intricate pattern of correlations. For any month h and any i such that $1 \leq i \leq p$, consider the panel which is interviewed in month $h+i$. There is an estimate from the rotation group which is in sample for the first time (disregard any groups used only for bounding purposes). This value is correlated with estimates from the same group for the previous $p-i$ months, but with nothing else. A second group is interviewed for the second time. Its estimate for month h is correlated with those for the prior $2p-i$ months. This pattern continues.

When the contributions and relationships of all the rotation groups in this panel have been sorted, one must bring in those from the other panels. Because each panel is interviewed in a different month, the corresponding covariances may be different. The entire process, although balanced and well-structured, is more intricate. This fact is reflected in the variance formulae for the generalized composite estimators of level and change. We plan to address these in a forthcoming report.

REFERENCES

Breau, P. and Ernst, L. R. (1983). Alternative Estimators to the Current Composite Estimator, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

Cantwell, P. (1988). Variance Formulae for the Generalized Composite Estimator Under a Balanced One-Level Rotation Plan, Statistical Research Division Research Report Series, No. 88-26, Bureau of the Census, Washington, D.C.

Nelson, D., McMillen, D. and Kasprzyk, D. (1984). An Overview of the Survey of Income and Program Participation, SIPP Working Paper Series, No. 8401, Bureau of the Census, Washington, D.C.

Wolter, K. M. (1979). Composite Estimation in Finite Populations, *Journal of the American Statistical Association*, 74, 604-613.

U. S. Department of Justice, Bureau of Justice Statistics (1988). *Criminal Victimization in the United States, 1986*, A National Crime Survey Report: NCJ-111456, Washington, D.C.

APPENDIX

Design Layout of Estimates For Several Months From p Panels, $p = 4$

MONTH ↓	PANELS → 1	2	3	4
1	$x_{1,4}$			
2	$x_{2,3}$	$x_{2,4}$		
3	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	
4	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$
5	$x_{5,4}$	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$
6	$x_{6,3}$	$x_{6,4}$	$x_{6,1}$	$x_{6,2}$
7	$x_{7,2}$	$x_{7,3}$	$x_{7,4}$	$x_{7,1}$
8	$x_{8,1}$	$x_{8,2}$	$x_{8,3}$	$x_{8,4}$
9	$x_{9,4}$	$x_{9,1}$	$x_{9,2}$	$x_{9,3}$
10	$x_{10,3}$	$x_{10,4}$	$x_{10,1}$	$x_{10,2}$
11	$x_{11,2}$	$x_{11,3}$	$x_{11,4}$	$x_{11,1}$
12	$x_{12,1}$	$x_{12,2}$	$x_{12,3}$	$x_{12,4}$
13	$x_{13,4}$	$x_{13,1}$	$x_{13,2}$	$x_{13,3}$
14	$x_{14,3}$	$x_{14,4}$	$x_{14,1}$	$x_{14,2}$
15	$x_{15,2}$	$x_{15,3}$	$x_{15,4}$	$x_{15,1}$
16	$x_{16,1}$	$x_{16,2}$	$x_{16,3}$	$x_{16,4}$
⋮	⋮	⋮	⋮	⋮

Note: $x_{h,i}$ denotes the estimate of "monthly" level for month h from the panel which is interviewed in month $h+i$. Solid horizontal lines separate estimates which are obtained in different interviews.

When $p = 4$:

$$J = \begin{bmatrix} 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0 \end{bmatrix}, \qquad J^2 = \begin{bmatrix} 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0 \end{bmatrix}, \qquad J^3 = \begin{bmatrix} 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0 \end{bmatrix}$$

$$J^4 = I, \qquad J^5 = J, \qquad J^6 = J^2, \qquad \ldots$$

$$D = \begin{bmatrix} d_1&0&0&0\\ 0&d_2&0&0\\ 0&0&d_3&0\\ 0&0&0&d_4 \end{bmatrix}, \qquad R_r = \begin{bmatrix} \rho_{r,1}&0&0&0\\ 0&\rho_{r,2}&0&0\\ 0&0&\rho_{r,3}&0\\ 0&0&0&\rho_{r,4} \end{bmatrix}, \quad r \geq 0$$

$$\mathrm{Cov}(x_h, x_{h-1}) = \sigma^2 D R_1 J D = \sigma^2 \begin{bmatrix} 0 & \rho_{1,1}d_1d_2 & 0 & 0\\ 0 & 0 & \rho_{1,2}d_2d_3 & 0\\ 0 & 0 & 0 & \rho_{1,3}d_3d_4\\ \rho_{1,4}d_4d_1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathrm{Cov}(x_h, x_{h-2}) = \sigma^2 D R_2 J^2 D = \sigma^2 \begin{bmatrix} 0 & 0 & \rho_{2,1}d_1d_3 & 0\\ 0 & 0 & 0 & \rho_{2,2}d_2d_4\\ \rho_{2,3}d_3d_1 & 0 & 0 & 0\\ 0 & \rho_{2,4}d_4d_2 & 0 & 0 \end{bmatrix}$$

. . .