On reuse of clusters in repeated studies

Stanislav Kolenikov and Gustavo Angeles
University of Missouri, Columbia, and University of North Carolina, Chapel Hill

Abstract

Suppose data for a survey with multi-stage design is to be collected in two periods of time. This paper assesses the relative merits of keeping the same clusters in the sample vs. sampling new clusters, under different statistical (correlation between clusters and over time) and logistical (costs of survey) scenarios. The design effect of re-using the same clusters from the master sample over time is of the form $1 - A\rho\pi/n$ where ρ is intertemporal correlation of the cluster totals, n is the number of clusters, π is the proportion of clusters retained from the previous round, and $A > 0$ is a fixed constant. As long as the efficiency gains appear to be minor, the value of the designs that reuse the clusters comes from the logistical (cost of the survey) considerations. Empirical demonstration that uses Demographic and Health Survey (DHS) data for Bangladesh, 1996 and 2000, is provided.

1. Introduction

A change in the population characteristic is often of interest to researchers and policymakers for the purposes of assessing the dynamics of population change, the effectiveness of economic or population health measures, and other research questions. At the level of a country, the data sources routinely used to address those questions are large complex surveys. When the survey is repeated over time, a new dimension of the sampling error component of the total survey error is due to the patterns of repeated observations, or sampling units. Non-sampling components of the error, such as (potentially informative) sample attrition, or conditioning (time in the sample) effects, are beyond the scope of this paper.

The literature about sampling on multiple occasion goes back to Jessen (1942), who considered single stage surveys at two occasions; Yates (1949), who extended this work into multiple occasions assuming that observations for each unit followed a stationary AR(1) process, and Patterson (1950) who studied a single stage survey on several occasions and estimation of the means for each occasion. At about that time, one of the largest regularly conducted US data collection efforts, the Current Population Survey, was conceived, that employed a 4-8-4 rotating design (Eckler 1955, Rao & Graham 1964, Binder & Hidiroglou 1988, U.S. Census Bureau 2002). Singh (1968) considered multi-stage designs for sampling on several occasions and discussed how the fractions of the earlier samples should be used in the later occasions, with application to an agricultural survey subjected to heavy seasonal variations. The area of repeated survey designs has recently achieved a new wave of interest from the natural resources research where a need arises to assess changes in forestation or agricultural health. In this area, the most popular designs are variations of the sampling with partial replacement scheme where each wave contains elements from all previous waves, as well as newly sampled units (Scott 1998, Fuller 1999, McDonald 2003).

The efficiency of estimating the change in population totals (means) implies that with observations positively correlated over time, the sampling designs allowing for extensive overlap of the sampling/observation units between waves provide more efficient estimates. Another practical consideration is the implementation cost (Groves 1989). An example where costs, potentially varying between units, play a major role in the sample design is Neyman-Tchuprow optimal allocation design (Neyman 1938). Our interest lies in the costs of repeated surveys: it can be argued that the costs of interviewing *individuals* increase with time (primarily, due to the costs of locating the household), while the cost of revisiting a *cluster* may be lower (the maps and population counts are already available; the cooperation with the community leaders and/or individual respondents has already been established; etc.).

This paper was motivated by the design of Demographic and Health Surveys[1], a U.S. Agency of International Development sponsored project that collects the family planning, maternal health, child survival, HIV/AIDS and other health information on over 70 developing countries. The surveys are highly standardized (subject to translation of the instruments into the country home languages). The sampling design includes stratification (by region and urbanicity) and clustering (by settlements). Typical sample sizes vary between 5,000 and 30,000 households. The clusters are revisited about every 5 years. A large period of time between consecutive interviews makes it impractical to locate the households interviewed previously, and new samples are taken at each of the locations. Thus there is a considerable overlap in the first stage sampling between time periods, while the second stage samples are taken independently. We shall refer to such designs as *cluster-panel designs*.

[1] See http://www.measuredhs.com.

2. Repeated surveys in simple settings

Suppose a SRSWR survey with sample size n is conducted on the population of size N.[2] The survey designer can control the amount of overlap between the two samples over time. Suppose a fraction π of units are the same in two periods of time, so that the first $(1-\pi)n$ observations are those taken at time $t=1$ and abandoned after that; then the next πn observations are those taken at both $t=1$ and $t=2$ by design; and the remaining $(1-\pi)n$ observations are those taken at time $t=2$ only by sampling with replacement independently from those retained from the previous period. Then two possible estimators of the mean change $\Delta = \bar{Y}_2 - \bar{Y}_1$ are *elementary estimate*

$$\hat{\delta} = \bar{y}_2 - \bar{y}_1 \tag{1}$$

and a *(one-step) composite estimate*

$$\hat{\delta}_\alpha = \frac{1}{n}\Big[(1+\pi\alpha)\Big(-\sum_{i=1}^{(1-\pi)n} y_{1i} + \sum_{i=n+1}^{(2-\pi)n} y_{2i}\Big) + (1-(1-\pi)\alpha)\sum_{i=(1-\pi)n+1}^{n} (y_{2i} - y_{1i})\Big] \tag{2}$$

where α controls the relative weights given to the panel subsample and the independent subsample. Let us assume that the variances of $Y_{1,i}$ and Y_2 are the same, $S_1^2 = S_2^2 = S^2$, and the intertemporal correlation is defined as

$$\frac{1}{N-1}\sum_{j=1}^{N}(y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2) = \rho S^2 \tag{3}$$

Then the variances of the two estimates are:

$$\mathbb{V}[\hat{\delta}] = \frac{2(1-\pi\rho)S^2}{n}, \tag{4}$$

$$\mathbb{V}[\hat{\delta}_\alpha] = \frac{2S^2}{n}\big[(1+\pi\alpha)^2(1-\pi) \tag{5}$$

$$+ (1-(1-\pi)\alpha)^2\pi(1-\rho)\big] \tag{6}$$

The optimal α that minimizes (6) is

$$\alpha^* = -\frac{\rho\pi}{2(1-\rho(1-\pi))} \tag{7}$$

and further variance minimization over the design parameter π gives

$$\mathbb{V}[\hat{\delta}_{\alpha^*}] = \frac{2S^2}{n}\frac{1-\rho}{1-\rho(1-\pi)} \tag{8}$$

Both (4) and (8) achieve their minima when $\pi = 1$, i.e., in the panel setting, and the two estimators coincide achieving variance $2(1-\rho)S^2/n$.

Let us now bring the cost considerations into account. Suppose the unit cost of observing the unit at the first stage only is c_1, the unit cost of observing the unit at both stages is c_{12}, and the unit cost of observing the unit only in the second stage is c_2. Then the optimal design for the elementary estimator is derived from the following minimization problem:

$$\frac{2(1-\pi\rho)S^2}{n} \to \min_{n,\pi}$$

$$\text{s.t. } c_1(1-\pi)n + c_{12}\pi n + c_2(1-\pi)n \le C_0 \tag{9}$$

The inequality is in fact binding (the budget of the survey is fully used), and the minimization problem becomes

$$\frac{2(1-\pi\rho)(c_1 + c_2 + \pi[c_{12} - c_1 - c_2])}{C_0} \to \min_{\pi\in[0,1]} \tag{10}$$

If $c_{12} \ne c_1 + c_2$, this equation gives a parabola with the center at

$$\pi^* = \frac{c_{12} - c_1(1+\rho) - c_2(1+\rho)}{2\rho(c_{12} - c_1 - c_2)} = \frac{1}{2\rho} - \frac{1}{2\left[\frac{c_{12}}{c_1+c_2} - 1\right]} \tag{11}$$

Let us now consider three special cases. If $c_{12} < c_1 + c_2$ (sampling new units is more costly than revisiting them; arguably, the case of the cluster designs), $\pi^* > 1$ gives the location of the minimum, and hence the optimal design is the one with $\pi = 1$. If $c_{12} = c_1 + c_2$, there are neither extra costs nor extra savings associated with the panel mode of data collection. In this case the objective function is linearly decreasing with π, and the optimal solution to it is $\pi = 1$. If $c_{12} > c_1 + c_2$ (sampling new individuals is cheaper than tracking them; arguably, the most realistic case in panel studies), the objective function in (10) is a parabola with downward branches, and the minimization problem has a corner solution at either 1 or 0, depending on which of those two points is further from the maximum:

$$\pi = \begin{cases} 0, & c_1 + c_2 < (1-\rho)c_{12} \\ 1, & c_1 + c_2 > (1-\rho)c_{12} \end{cases} \tag{12}$$

When $c_1 + c_2 = (1-\rho)c_{12}$, both sample designs with $\pi = 0$ and $\pi = 1$ give the same variance.

For the one-step composite estimator of change, the analogous minimization problem is

$$\frac{2S^2}{C_0}\frac{(1-\rho)(c_1 + c_2 + \pi(c_{12} - c_1 - c_2))}{1-\rho(1-\pi)} \to \min_{\pi} \tag{13}$$

and it turns out to have the same optimality conditions as (10), since the optimal designs under the latter one are either independent sampling, or panel study, in which case both the elementary estimate and the composite estimate coincide.

[2] The capital letters denote the population quantities, and the lower case letters, the sample ones. Time periods are denoted by subscripts.

3. Repeated surveys with cluster sampling

Most large multistage surveys are collected using cluster sampling at certain stages where a groups of observation units, rather than units themselves, are sampled at the early stages of sampling, possibly with varying probabilities (say probability proportional to size). In DHS studies that are our main motivation, the clusters are settlements, such as villages in rural areas, or city districts in urban areas. For the analysis in this section, we consider a two-stage cluster equal probability of selection (epsem) design where the finite population corrections may be ignored. More complicated probability designs will produce conceptually the same results attenuated by weights and fpcs.

Let us denote the clusters by i, so that there are N clusters in the population and n clusters in the sample. Let us enumerate the observations within i-th cluster by j, so that $j = 1, \dots, M_i$ for i-th cluster in population, and $j = 1, \dots, m_i$ in the sample. j-th observation in i-th cluster is denoted as Y_{ij} in the population or y_{ij} in the sample. The totals and their estimates are denoted by $T[Y]$ and $t[y]$ in population and in the sample, respectively. The means per observation units are ratios of the corresponding totals of Y or y, and the totals of 1's, in the clusters or populations. The total variance of the response variable is S^2, the variance within i-th cluster, S^2_{wi}, and the variance between clusters, S^2_b.

Derivation of the variances of the totals and means can be found in a number of standard textbooks (Hansen, Hurwitz & Madow 1953, Kish 1965, Thompson 1992, Särndal, Swensson & Wretman 1992):

$$\mathbb{V}[t_{\cdot\cdot}] = N^2 \frac{1-f_I}{n} S^2_b + N/n \sum_{i=1}^{n} M_i^2 \frac{1-f_i}{m_i} S^2_{wi} \quad (14)$$

Suppose now that the survey is repeated over time, so that there are at least two waves of data. Denote the time by an upper index: $Y^{(t)}_{ij}$, $t = 1, 2, \dots$

The quantity of interest to the researcher would be the difference in population totals or, more often, averages per observation unit of characteristic Y:

$$D[\bar{Y}^{(2)} - \bar{Y}^{(1)}] = \frac{T[Y^{(2)}]}{T[1^{(2)}]} - \frac{T[Y^{(1)}]}{T[1^{(1)}]} \qquad (15)$$

For general ratio estimators of y/x, the estimator of (15), although biased in finite samples, is the difference of the corresponding ratio estimators:

$$\begin{aligned} d[\bar{y}^{(2)} - \bar{y}^{(1)}] &= \frac{t[y^{(2)}]}{t[x^{(2)}]} - \frac{t[y^{(1)}]}{t[x^{(1)}]} \\ &= d\big(t[y^{(2)}], t[y^{(1)}], t[x^{(2)}], t[x^{(1)}]\big) \quad (16) \end{aligned}$$

and its variance is

$$\begin{aligned} \mathbb{V}\big[d[y^{(2)} - y^{(1)}]\big] &= \mathbb{V}\Big[\frac{T[Y^{(2)}]}{T[X^{(2)}]}\Big] + \mathbb{V}\Big[\frac{T[Y^{(1)}]}{T[X^{(1)}]}\Big] \\ &\quad - 2\,\mathbb{C}\mathrm{ov}\Big[\frac{T[Y^{(2)}]}{T[X^{(2)}]}, \frac{T[Y^{(1)}]}{T[X^{(1)}]}\Big] \quad (17) \end{aligned}$$

In terms of the previous section, it corresponds to the elementary estimate of the change. The composite estimates do not seem to be frequently used in large surveys, as computing those estimates will either require supplying a new set of weights by the institution collecting the data, or estimating the intertemporal correlation coefficient by the user of the data. Any of those procedures will be specific to the difference being estimated, and will tend to be rather cumbersome.

The linear approximations for the first two terms of (17) can be found using (14) and linearization technique. If the samples in different periods are taken independently of one another, then the third term is zero. The case we are interested in, however, is when the clusters from the first sample are reused, at least partially, in the second sample.

In computing the linearization approximation to the last covariance term, it should be noted that for the cluster-panel designs the covariances across time can be simplified as follows:

$$\begin{aligned} &\mathbb{C}\mathrm{ov}\big[t[\xi^{(2)}], t[\zeta^{(1)}]\big] = \\ &= \mathbb{E}_{\mathrm{I}}\Big\{\mathbb{C}\mathrm{ov}_{\mathrm{II}}\big[t[\xi^{(2)}], t[\zeta^{(1)}] | \mathrm{I}\big]\Big\} \\ &+ \mathbb{C}\mathrm{ov}_{\mathrm{I}}\Big\{\mathbb{E}_{\mathrm{II}}\big[t[\xi^{(2)}] | \mathrm{I}\big], \mathbb{E}_{\mathrm{II}}\big[t[\zeta^{(1)}] | \mathrm{I}\big] \\ &= \frac{N^2 \pi n}{n^2}\, \mathbb{C}\mathrm{ov}\big[\Xi^{(2)}_{i\cdot}, \mathrm{Z}^{(1)}_{i\cdot}\big] \\ &= \frac{\pi N^2}{(N-1)n} \sum_{i=1}^{N} (\Xi^{(2)}_{i\cdot} - \bar{\Xi}^{(2)}_{i\cdot})(\mathrm{Z}^{(1)}_{i\cdot} - \bar{\mathrm{Z}}^{(1)}_{i\cdot})\Big\} \quad (18) \end{aligned}$$

where indices I and II represent the first and the second stages of the sampling, respectively. The first conditional covariance on the second line is zero, as long as sampling at the second stage is performed independently across waves.

In computing the means per unit, $x_{ij} = 1$. Let us also make a simplifying assumption that the design is fixed size, so $\mathbb{V}\big[t[1^{(t)}]\big] = 0$, $t = 1, 2, \dots$. Then the covariance term in (17) is comprised only of the covariances between y's in two time periods:

$$\begin{aligned} &\mathbb{C}\mathrm{ov}\Big[\frac{t[y^{(2)}]}{t[x^{(2)}]}, \frac{t[y^{(1)}]}{t[x^{(1)}]}\Big] \approx \frac{\mathbb{C}\mathrm{ov}\big[t[y^{(2)}], t[y^{(1)}]\big]}{T[X^{(1)}]T[X^{(2)}]} \approx \\ &\approx \frac{N^2 \pi\, \mathbb{C}\mathrm{ov}\big[Y^{(2)}_{i\cdot}, Y^{(1)}_{i\cdot}\big]}{(N\bar{M})^2 n} = \frac{\pi\, \mathbb{C}\mathrm{ov}\big[Y^{(2)}_{i\cdot}, Y^{(1)}_{i\cdot}\big]}{\bar{M}^2 n} \quad (19) \end{aligned}$$

Then the variance of the difference can be found as

$$\mathbb{V}\big[d[y^{(2)}-y^{(1)}]\big] \approx \frac{1}{N^2\bar{M}^2}\,\mathbb{V}\big[t[y^{(2)}]\big] + \frac{1}{N^2\bar{M}^2}\,\mathbb{V}\big[t[y^{(1)}]\big] - 2\frac{\pi\,\mathbb{C}\mathrm{ov}\big[Y_{i\cdot}^{(2)},Y_{i\cdot}^{(1)}\big]}{\bar{M}^2 n} \tag{20}$$

If the last covariance is positive (i.e., the clusters with higher values of Y in the first period continue to have higher values in the second period), then the re-use of clusters will be decreasing variance: the higher the proportion of reused clusters π, the lower the total variance (20). Again, if the sampling is performed independently in the two waves of data collection, the last term is zero. Setting it to zero also corresponds to the naïve estimator of the difference variance that does not account for the longitudinal nature of the data collection process. Thus the design effect of repeated sampling that compares the naïve estimate with the appropriate one is

$$DEFF_r = \frac{\mathbb{V}[\text{repeated design}]}{\mathbb{V}[\text{independent sampling}]} = 1 - 2\frac{\pi\,\mathbb{C}\mathrm{ov}\big[Y_{i\cdot}^{(2)},Y_{i\cdot}^{(1)}\big]}{n(\mathbb{V}\big[t[y^{(1)}]\big]+\mathbb{V}\big[t[y^{(2)}]\big])/N^2} \tag{21}$$

so the correction is in fact of the order $O(n^{-1})$, and the repeated sampling design effect is going to be small unless the number of clusters is small (say 20 or less), which is against the standard clustered design recommendation of having many clusters with few observations per cluster. The naïve variance estimator is conservative for positive $\mathbb{C}\mathrm{ov}\big[Y_{i\cdot}^{(2)},Y_{i\cdot}^{(1)}\big]$, and is consistent when $n \to \infty$.

4. Costs for repeated cluster samples

This section will analyze the cost efficiency of clustered samples when one wants to estimate the difference between two sample means from two different periods.

Some discussion of the costs of cluster sampling is given in Thompson (1992, Sec. 12.5), and more mathematical details are available in Hansen et. al. (1953, Vol. II, Sec. 6.11) with the variance formulas corrected for finite populations.

Let us assume the following cost structure: c_1^{I} is the cost of sampling and collecting the community data at time $t=1$ for the clusters that are used *in the first wave only*; c_1^{II} is the cost of sampling and interviewing an individual at time $t=1$; c_2^{I} is the cost of sampling a *new* cluster at time $t=2$; c_2^{II} is the cost of sampling and interviewing an individual at time $t=2$; c_{12}^{I} is the cost of sampling and collecting the data for clusters that have the data collected in both periods $t=1$ and $t=2$.

Let the population consist of N clusters in both time periods, and each cluster consist of M individuals. Let the number of clusters used in *only* the first time period be n_1, *only* in the second period, n_2, and the number of clusters used in both waves, n_0. Let the number of units sampled in each cluster be m_1 in the first wave and m_2 in the second wave. Then the total variable cost of the survey is

$$C_0 = c_1^{\mathrm{I}}n_1+c_{12}^{\mathrm{I}}n_0+c_2^{\mathrm{I}}n_2+c_1^{\mathrm{II}}(n_1+n_0)m_1+c_2^{\mathrm{II}}(n_2+n_0)m_2 \tag{22}$$

The sample designer wishes to minimize the variance of the elementary difference estimator (17):

$$\{n_0,n_1,n_2,m_1,m_2\} = \arg\min \mathbb{V}\Big[\frac{t[y^{(2)}]}{t[1^{(2)}]} - \frac{t[y^{(1)}]}{t[1^{(1)}]}\Big] \tag{23}$$

Note that this is objective function focuses solely on the difference between the two sample means, while in the practical situation, the design should also allow for efficient estimation of the contemporary means.

Note that the design is of the fixed size, so $\mathbb{V}\big[t[1^{(t)}]\big]=0$, $t=1,2$. From the results in two preceding sections, the variance of (23) is

$$\begin{aligned}\mathbb{V}\Big[\frac{t[y^{(2)}]}{t[1^{(2)}]} - \frac{t[y^{(1)}]}{t[1^{(1)}]}\Big] &= \frac{N-(n_1+n_0)}{(n_1+n_0)NM^2}S_{1b}^2 + \frac{1}{NM}\frac{M-m_1}{m_1}\bar{S}_{1w}^2 \\ &+ \frac{N-(n_2+n_0)}{(n_2+n_0)NM^2}S_{2b}^2 + \frac{1}{NM}\frac{M-m_2}{m_2}\bar{S}_{2w}^2 \\ &-2\frac{n_0\rho^{\mathrm{I}}S_{1b}S_{2b}}{(n_1+n_0)(n_2+n_0)M^2}\end{aligned} \tag{24}$$

where the variance (14) was used for the first two terms, and

$$\rho^{\mathrm{I}} = \frac{\sum_{i=1}^{N}\big(Y_{i\cdot}^{(2)}-\bar{Y}_{i\cdot}^{(2)}\big)\big(Y_{i\cdot}^{(1)}-\bar{Y}_{i\cdot}^{(1)}\big)}{(N-1)M^2S_{1b}S_{2b}} \tag{25}$$

is the intertemporal correlation of the cluster totals.

The minimization constraints are:

$$c_1^{\mathrm{I}}n_1+c_{12}^{\mathrm{I}}n_0+c_2^{\mathrm{I}}n_2 + c_1^{\mathrm{II}}(n_1+n_0)m_1+c_2^{\mathrm{II}}(n_2+n_0)m_2 \le C_0, \tag{26}$$

$$n_0 \ge 0, \quad n_1 \ge 0, \quad n_2 \ge 0 \tag{27}$$

and with the corresponding Lagrange multipliers λ, λ_0, λ_1, λ_2, the Lagrangian function $L(n_0,n_1,n_2,m_1,m_2;\lambda,\lambda_0,\lambda_1,\lambda_2)$ can be written down as a combination of (24), (26) and (27). For details and derivations, see Kolenikov & Angeles (2005).

The necessary conditions for this minimization problem will be considered for three cases of the greatest interest.

Case 1: independent sampling $n_0 = 0$, $\lambda_1 = \lambda_2 = 0$, n_1, $n_2 > 0$. No common clusters are sampled in two periods of time; all of the sampling is performed independently.

Case 2: cluster-panel design $\lambda_0 = 0$, $n_1 = n_2 = 0$. All of the clusters sampled in the first period are reused again in the second period.

Case 3: mixed design $\lambda_0 = \lambda_1 = \lambda_0 = 0$, n_0, n_1, $n_2 > 0$. At each time period, the sample contains both clusters common to the two observation periods, and independent wave-specific clusters.

Other cases, such as $n_1, n_0 > 0$, $n_2 = 0$, will not arise because of the symmetry of the problem with respect to time $t = 1, 2$.

4.1 Independent sampling

If the optimal design is such that the samples are taken independently in two periods of time, so that $n_0 = 0$, then also the Lagrange multipliers for constraints on n_1 and n_2 are zero. Substituting this to the necessary conditions, one obtains the following set of equations:

$$n_1 = \frac{S_{1b}^2 N m_1^2 c_1^{\mathrm{II}}}{M^2 \bar{S}_{1w}^2 (c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)} \tag{28}$$

$$n_2 = \frac{S_{2b}^2 N m_2^2 c_2^{\mathrm{II}}}{M^2 \bar{S}_{2w}^2 (c_2^{\mathrm{I}} + c_2^{\mathrm{II}} m_2)} \tag{29}$$

$$c_1^{\mathrm{I}} n_1 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}} n_1 m_1 + c_2^{\mathrm{II}} n_2 m_2 = C_0, \tag{30}$$

$$\frac{\bar{S}_{1w}^2}{m_1^2 c_1^{\mathrm{II}} n_1} = \frac{\bar{S}_{2w}^2}{m_2^2 c_2^{\mathrm{II}} n_2} \tag{31}$$

Simpler answers can be obtained assuming *equal conditions*, i.e., that the costs and variances do not change between the two periods:

$$S_{1b}^2 = S_{2b}^2 = S_b^2,\ \bar{S}_{1w}^2 = \bar{S}_{2w}^2 = \bar{S}_w^2,$$
$$c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = c^{\mathrm{I}},\ c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = c^{\mathrm{II}} \tag{32}$$

Then the number of clusters and cluster size are

$$m = M\sqrt{\frac{C_0 \bar{S}_w^2}{2Nc^{\mathrm{II}} S_b^2}}, n = \frac{C_0}{2\left[c^{\mathrm{I}} + M\sqrt{C_0 c^{\mathrm{II}} \bar{S}_w^2 / 2N S_b^2}\right]} \tag{33}$$

so both m and n increase as ${C_0}^{1/2}$ for large surveys (although $n \propto C_0$ for smaller ones).

From (24), the variance of the difference estimator is

$$\mathbb{V}_{e,i}[d] \approx \frac{4S_b^2\left[c^{\mathrm{I}} + \sqrt{M^2 C_0 c^{\mathrm{II}} \bar{S}_w^2 / 2N S_b^2}\right]}{C_0 M^2}(1 - \rho^{\mathrm{I}})$$
$$+ 2\sqrt{\frac{2c^{\mathrm{II}} S_b^2 \bar{S}_w^2}{N C_0 M^2}} \tag{34}$$

where the (conservative) approximation is made by setting the finite population corrections to zero (i.e., $n \ll N$, $m \ll M$), and the subindex e, i stands for "equal conditions — independent samples".

4.2 Cluster-panel design

If the design with $n_1 = n_2 = 0$ is optimal, then $\lambda_0 = 0$, and the system of necessary conditions leads to

$$m_1 = \sqrt{C_0/U}, \quad m_2 = \kappa\sqrt{C_0/U},$$
$$n_0 = \frac{C_0}{c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2}, \quad \kappa = \sqrt{\frac{\bar{S}_{2w}^2 c_1^{\mathrm{II}}}{\bar{S}_{1w}^2 c_2^{\mathrm{II}}}},$$
$$U = \frac{(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}) N\kappa (c_1^{\mathrm{II}} c_2^{\mathrm{II}})^{\frac{1}{2}}}{M^2 \bar{S}_{1w} \bar{S}_{2w}} \tag{35}$$

Again, the number of units per cluster increases with the budget as $\sqrt{C_0}$, and the number of clusters sampled increases as C_0 for small surveys, and as $\sqrt{C_0}$, for large ones. The variance of the difference estimator can now be found simplifying (24) as

$$\mathbb{V}[d] \approx \frac{U^{\frac{1}{2}}}{N{C_0}^{\frac{1}{2}}}(\bar{S}_{1w}^2 + \bar{S}_{2w}^2)$$
$$+ \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{C_0 M^2}\left(c_{12}^{\mathrm{I}} + \frac{{C_0}^{\frac{1}{2}}}{U^{\frac{1}{2}}}(c_1^{\mathrm{II}} + \kappa c_2^{\mathrm{II}})\right) \tag{36}$$

Under equal conditions assumption (32),

$$\kappa = 1, \quad U = \frac{2S_b^2(1 - \rho^{\mathrm{I}}) N c^{\mathrm{II}}}{M^2 \bar{S}_w^2},$$
$$\mathbb{V}_{e,p}[d] \approx 2\sqrt{\frac{2S_b^2 \bar{S}_w^2 (1 - \rho^{\mathrm{I}}) c^{\mathrm{II}}}{M^2 N C_0}}$$
$$+ \frac{2S_b^2(1 - \rho^{\mathrm{I}})}{C_0 M^2}\left(c_{12}^{\mathrm{I}} + 2\sqrt{\frac{M^2 \bar{S}_w^2 C_0 c^{\mathrm{II}}}{2S_b^2(1 - \rho^{\mathrm{I}}) N}}\right) \tag{37}$$

where the subindex e, p stands for "equal conditions — panel clusters".

4.3 Comparison of the independent and panel-cluster designs

The difference of two variances (34) and (37) is

$$\mathbb{V}_{e,p}[d] - \mathbb{V}_{e,i}[d] = \frac{2S_b^2(1 - \rho^{\mathrm{I}})}{C_0 M^2}(c_{12}^{\mathrm{I}} - 2c^{\mathrm{I}})$$
$$-2\sqrt{\frac{2\bar{S}_w^2 S_b^2 c^{\mathrm{II}}}{M^2 N C_0}}\left(1 - \sqrt{1 - \rho^{\mathrm{I}}}\right)^2 \tag{38}$$

The last term is always negative, and the cluster-panel design is guaranteed to be more efficient when $c_{12}^{\mathrm{I}} \le 2c^{\mathrm{I}}$, i.e., when revising the clusters indeed provides cost savings. Note also that as $\rho^{\mathrm{I}} \to 1$ (i.e., the characteristic is persistent and does not change much between rounds), the first term goes to zero, while the second term converes to a fixed negative quantity, so the cluster-panel design is more efficient even when the re-use of clusters is more expensive than sampling new clusters. Also, the second term decreases slower than the first one with the size of the survey, and the cluster-panel design may be more variance-efficient even when it is slightly more expensive to collect the data in that manner:

$$c^{\mathrm{I}}_{12} < 2c^{\mathrm{I}} + \sqrt{\frac{2\bar{S}^2_w M^2 c^{\mathrm{II}} C_0}{S^2_b N(1-\rho^{\mathrm{I}})^2}} \bigl(1 - \sqrt{1-\rho^{\mathrm{I}}}\bigr)^2 \qquad (39)$$

This preference for a panel-cluster designs will be stronger for larger surveys with higher total budget C_0.

4.4 Intermediate case

The design satisfying the first order that has all of $n_0, n_1, n_2 > 0$ is difficult to characterize. Under the equal condition assumption (32), the results are:

$$\frac{S^2_b(n+n_0-2n_0\rho^{\mathrm{I}})}{M^2(n+n_0)^2} = \frac{\bar{S}^2_w(c^{\mathrm{I}}+c^{\mathrm{II}}m)}{Nm^2c^{\mathrm{II}}}, \qquad (40)$$

$$\frac{2S^2_b(n+n_0+\rho^{\mathrm{I}}(n-n_0))}{M^2(n+n_0)^2} = \frac{\bar{S}^2_w}{Nm^2c^{\mathrm{II}}}(c^{\mathrm{I}}_{12}+2c^{\mathrm{II}}m) \qquad (41)$$

$$2c^{\mathrm{I}}n + c^{\mathrm{I}}_{12}n_0 + 2c^{\mathrm{II}}(n+n_0)m = C_0 \qquad (42)$$

Introducing

$$\nu = \frac{n_0}{n}, \quad \pi^{-1} = 1+\nu^{-1} \qquad (43)$$

and dividing the first equation by the second one, one gets

$$m = \frac{A+B\nu}{C+D\nu},$$
$$A = c^{\mathrm{I}}(1+\rho^{\mathrm{I}}) - c^{\mathrm{I}}_{12}, \quad B = c^{\mathrm{I}}(1-\rho^{\mathrm{I}}) - c^{\mathrm{I}}_{12}(1-2\rho^{\mathrm{I}}),$$
$$C = c^{\mathrm{II}}(1-\rho^{\mathrm{I}}), \quad D = c^{\mathrm{II}}(1-3\rho^{\mathrm{I}}) \qquad (44)$$

Analyzing the range of ρ^{I} and ν where m from (44) can be positive, the following existence result can be established: there exist characteristics values of the intertemporal correlations

$$\exists\,\rho_1 : \frac{c^{\mathrm{I}}_{12} - c^{\mathrm{I}}}{2c^{\mathrm{I}}_{12} - c^{\mathrm{I}}} < \rho_1 < \frac{1}{3}$$
$$\exists\,\rho_2 : \frac{c^{\mathrm{I}}_{12} - c^{\mathrm{I}}}{c^{\mathrm{I}}} < \rho_2 < 1 \qquad (45)$$

so that in the range $\rho_1 < \rho^{\mathrm{I}} < \rho_2$, the first order conditions can be satisfied.

For $\rho^{\mathrm{I}} < \rho_1$, the necessary conditions of the Lagrange multiplier problem are incompatible with one another, and hence the optimal design is one of the independent sampling or panel-cluster designs. Also for $\rho^{\mathrm{I}} > 1/2$, the optimal design has $\nu < 1$, i.e., $n_0 < n$, which seems counterintuitive. This tends to indicate that this design may correspond to the local maximum rather than the local minimum of the variance.

4.5 Numerical illustration

A short numerical example illustrates the above formulae and results. Consider the population defined the following set of parameters:

$$N = 2000, \quad M = 200, \quad S_b = 1.5, \quad S_w = 1,$$
$$c^{\mathrm{I}} = 1, \quad c^{\mathrm{I}}_{12} = 1.7, \quad c^{\mathrm{II}} = 0.25, \quad C_0 = 500$$

The number of clusters sampled is shown on Fig. 1. The characteristic intertemporal correlations relevant for the mixed design are: $\rho_1 = 0.292, \rho_2 = 0.997$. The design optimal for estimation of the mean on half budget, or the elementary estimate (the independent clusters design) has 10 clusters with 96 units per cluster (the solid horizontal line on the plot).

As ρ^{I} increases, the cluster-panel design tends to sacrifice n_0 in favor of m, so that the cluster means and differences are more accurately estimated. In the limit of $\rho^{\mathrm{I}} \to 1$, it suffices to have 1 cluster to estimate the change; however $n_0 = 5$ clusters are sampled for $\rho^{\mathrm{I}} > 0.72$ as long as the optimal cluster size hits the restriction $m \le M$. The mixed design has only slightly varying number of clusters ($n + n_0$ fluctuates between 27 and 31), with the changes in ρ^{I} influencing allocation between the independent and the panel portions of the clusters.

The total sample sizes are between 891 and 918 for mixed designs, and between 970 and 985 for cluster-panel designs.

Finally, the most important plot is that of the variance of the elementary estimate given on Fig. 2. It clearly shows the advantage of the cluster-panel design over other options. The design effects follow a similar pattern, with the reference line of DEFF for the independent cluster design being 14.3, and the DEFF for the cluster-panel design falling from that figure down to 0.12.

As an overall conclusion of this small numerical illustration, it appears that the cluster-panel design is the most variance-efficient for a given cost.

4.6 Remarks

One of the assumptions used in deriving the above results was that (the composition of) the population itself does not change: no units leave the population, and no new units appear. This is quite a restrictive assumption for many practical situations, and the sample designer might still want to include new clusters into the second wave of data collection if the population has changed between the two waves. Then the new clusters can be joined into a separate stratum, and a clustered sample can be taken from that stratum. Also, the dynamic measurement effects such as conditioning and time in sample lead to rotation bias, so it might be beneficial to provide at least some rotation of the PSUs. For the DHS studies, in particular, the first argument (coverage) is likely to be more important than the second one (time in sample) due to a substantial time between the waves of the survey (about 5 years).

5. Empirical illustration

The empirical illustration of the differences in designs is carried out with DHS data from Bangladesh, 1996 and 2000 data. Table 1 lists the results for different designs, and for two different measures, one of which (contraceptive use in married women) has a lot of individual level variability with little between cluster variability, and thus moderate design effects, and the other one (access to tap water) has extremely strong patterns among communities. Δ-naïve estimator of difference is the one that *does not take* into account the same clusters. Δ-design is the estimator that *does take* into account that the same clusters were used in two years. The line "Longitudinal effect" is the difference in variances of the Δ-naïve and Δ-design estimators. There does not seem to be much difference between the two, as it is within 3% for both measures. This is in accordance with the above theoretical argument that the design effect is close to 1 for a study with a large number of clusters. Also, there is relatively modest overlap in clusters: out of 313 clusters in the first study, only 137 were used for the consecutive study, and there were 204 new clusters.

Table 1: Differences in variance estimates in different design specifications.

Item	Estimate	S.e.	DEFF
Contraceptive use			
1996	49.24%	1.098%	4.072
2000	53.77%	0.941%	3.466
Δ-naïve	4.53%	1.446%	3.789
Δ-design	4.53%	1.431%	3.714
Estimated longitudinal effect			1.020
Access to tap water			
1996	5.24%	0.946%	85.77
2000	6.17%	1.039%	101.24
Δ-naïve	0.928%	1.422%	95.91
Δ-design	0.928%	1.405%	93.56
Estimated longitudinal effect			0.976

Source: Bangladesh DHS, 1996 and 2000.

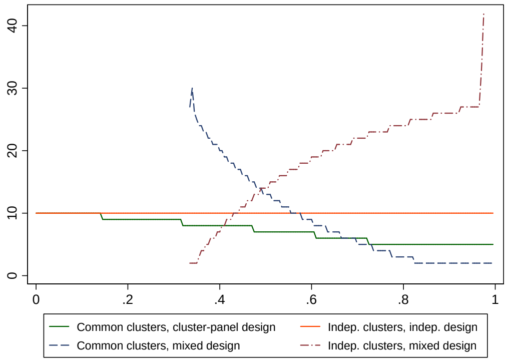

Figure 1: Number of clusters as a function of ρ^{I}.

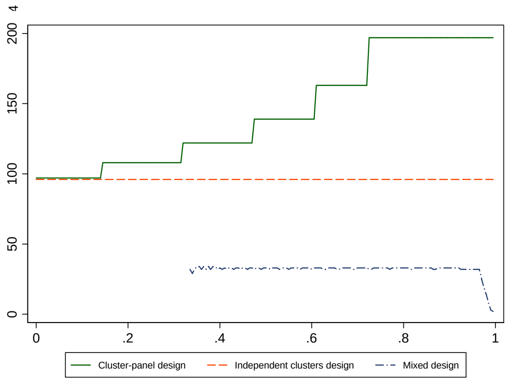

Figure 2: $\mathbb{V}\,d[\bar{y}^{(2)} - \bar{y}^{(1)}]$ as a function of ρ^{I}.

5.1 Outline of Stata code

The substantial part of Stata code (version 8) does the following:

1. creates `year` variable for two time periods;

2. sets the survey data configuration appropriately for the Δ-naïve estimator:
   ```
   egen psuXyear = group(psu year)
   svyset [pw=weight] , psu(psuXyear)
   ```

3. the difference of interest can be obtained in two possible ways: (i) as the difference in means,
   ```
   svymean depvar, by(year) deff
   lincom depvar[2000] - depvar[1996],
   deff
   ```

or (ii) as the regression coefficient of a dummy variable:

```
xi :  svyreg depvar i.year, deff
```

4. sets the survey data configuration appropriately for the Δ-design estimator:
   `svyset [pw=weight] , psu(psu)`

5. repeats step 3 for this design setting.

6. The longitudinal design effect is finally obtained as the ratio of two estimators of the variance.

6. Conclusions

This paper has analyzed the effect of re-using the clusters in repeated clustered surveys. The two main results of the paper are (i) that the design effect of correctly specifying the repeated use of clusters vs. assuming the two samples were taken independently are of the order $O(\rho\pi/n)$ where n is the number of clusters, ρ is the intertemporal correlation of cluster means, and π is the degree of overlap between two consecutive samples; and (ii) that for a given budget of the survey, the the designs that reuse the master sample clusters (referred to as cluster-panel designs) are more variance efficient for difference estimations than the design where the samples are taken anew. The difference in variances depends on the intertemporal correlation and the size of the survey. The considerations in favor of the panel-cluster designs come from the logistical side rather than from variance considerations, and a sample designer who knows that the characteristic of interest is going to have some degree of persistence over time will choose the cluster-panel design, unless it is known that the cost of re-visiting the first wave clusters are prohibitively high.

Acknowledgements

The authors are grateful to Chris Skinner and William Kalsbeek for their helpful comments. Nash Herndon provided editorial support. Financial support was provided by U.S. Agency for International Development through the MEASURE Evaluation project of Carolina Population Center, University of North Carolina at Chapel Hill, under the terms of Cooperative Agreement GPO-A-00-03-00003-00.

7. References

Binder, D. A. and Hidiroglou, M. A. (1988), Sampling in time, *in* P. R. Krishnaiah and C. R. Rao, eds, 'Handbook of Statistics', Vol. 6, North Holland, Amsterdam, pp. 187–211.

Eckler, A. R. (1955), 'Rotation sampling', *Annals of Mathematical Statistics* **26**(4), 664–685.

Fuller, W. A. (1999), 'Environmental surveys over time', *Journal of Agricultural, Biological and Environmental Statistics* **4**(4), 331–345.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, John Wiley & Sons, New York.

Hansen, M., Hurwitz, W. N. and Madow, W. G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons, New York.

Jessen, R. J. (1942), 'Statistical investigaion of a farm survey for obtaining farm facts', *Iowa Agricultural Station Research Bulletin* **304**, 54–59.

Kish, L. (1965), *Survey Sampling*, John Wiley and Sons, New York.

Kolenikov, S. and Angeles, G. (2005), On reuse of clusters in repeated studies, Working paper, Carolina Population Center, University of North Carolina, Chapel Hill.

McDonald, T. L. (2003), 'Review of environmental monitoring methods: Survey designs', *Environmental Monitoring and Assessment* **85**, 277–292.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *The Journal of the American Statistical Association* **33**, 101–116.

Patterson, H. D. (1950), 'Sampling on successive occasions with partial replacement of units', *Journal of the Royal Statistical Society, Series B* **12**(2), 241–255.

Rao, J. N. K. and Graham, J. E. (1964), 'Rotation designs for sampling on repeated occasions', *Journal of the American Statistical Association* **59**(306), 492–509.

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

Scott, C. T. (1998), 'Sampling methods for estimating change in forest resources', *Ecological Applications* **8**(2), 228–233.

Singh, D. (1968), 'Estimates in successive sampling using a multi-stage design', *Journal of the American Statistical Association* **63**(321), 99–112.

Thompson, S. K. (1992), *Sampling*, John Wiley and Sons, New York.

U.S. Census Bureau (2002), Current population survey: Design and methodology, Technical Paper 63RV, U.S. Census Bureau, Washington, DC. http://www.census.gov/prod/2002pubs/tp63rv.pdf.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*, Charles Griffin, London.