On reuse of clusters in repeated studies

Stanislav Kolenikov[*] Gustavo Angeles[†]

November 6, 2005

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 demonstrations that use Demographic and Health Survey (DHS) data for Bangladesh, 1996 and 2000, and a Monte Carlo simulation, are provided.

1 Introduction

In many situations, researchers or policymakers need to obtain information on characteristics of population on multiple occasions, or several times, to study the change in those characteristics and monitor the trends and current conditions. As long as complex design surveys are the most frequently used source of such information for a large population, such as a country as a whole, there is a need to study the properties of complex sampling designs when the survey is conducted in several points in time. Besides the standard considerations in single-shot surveys, such as stratification and clustering, and their effect on the cost and the estimate variances, there are additional concerns such as the desirable overlap among consecutive samples, (informative) sample attrition, conditioning effects, and other dynamic factors in the total survey error.

For the purposes of estimation of change from repeated surveys, it is often desirable to have high temporal correlation of the observation units which can be

[*] Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100. Email: kolenikovs@missouri.edu.

[†] CB# 7445, Rosenau Hall, University of North Carolina, Chapel Hill, NC 27599-7445. Email: gangeles@email.unc.edu.

achieved by administering the survey to the same sampling and/or observation units. The most popular survey design formats appear to be the longitudinal, or panel, surveys, where the same units (individuals, households) are revisited for several periods, potentially indefinitely many periods (studies like Panel Study of Income Dynamics (PSID), British Household Panel Study (BHPS) and others[1]); rotating panel where the units are recruited into the sample for a few periods, then rotated out of the sample, and interviewed again at a later time (Current Population Survey, CPS), or studies where only primary units are retained from one period to the next (Demographic and Health Survey, DHS).

The earliest paper to consider sampling on multiple occasions appears to be Jessen (1942), which considered single stage surveys at two occasions. This work was extended to multiple occasions by Yates (1949) who assumed that observations for each unit followed a stationary AR(1) process. Patterson (1950) studied a single stage survey on several occasions and estimation of the means for each occasion, assuming arbitrary correlation structure. He derived the efficient estimates of the contemporary means that use information from other periods (composite estimators), and efficient estimates of the change between the two last periods. 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, and applied his results to the analysis of an agricultural survey that is subjected to heavy seasonal variations.

The literature since the 1950s has mostly been devoted to the studies of rotating designs (Eckler 1955, Rao & Graham 1964, Binder & Hidiroglou 1988). Those designs compromise between accuracy in change estimation of the panel designs where the same units are interviewed over time, and deterioration of the sample quality as a consequence of sample attrition due to response burden, changes in the measurement process (time in sample, or conditioning effects, as the respondent behavior changes due to the fact that they are studied), and due to loss of population coverage due to underlying population changes.

A lot of research related to the rotating designs is devoted to Current Population Survey that uses 4-8-4 rotating design (Binder & Hidiroglou 1988, U.S. Census Bureau 2002) where units (households) are interviewed each month for 4 months, then receive a rest from the response burden for 8 months, and then interviewed for 4 months again. At each month, there are 8 cohorts of households recruited at different points in time, and this design achieves an overlap of 3/4 of the sample for estimates of the monthly changes, and 1/2 for estimates of annual changes.

Another large area of the rotating design applications is natural resource research, where a need arises to assess changes in forestation or agricultural health (McDonald 2003, Fuller 1999). The designs that have become popular in this area are sampling with partial replacement (SPR) designs (Scott 1998). In those designs, the plots established at each of the previous occasions are sampled and remeasured at the current occasion; in other words, at every time

[1] A compendium of information on the longitudinal studies can be found at Institute for Social and Economics Research web site, `http://iser.essex.ac.uk/ulsc/keeptrack/index.php`.

t there are units that were first sampled at each of the previous times $t-1$, $t-2$, $\ldots$ and the newly added plots. Such designs allow for efficient estimation of both the current characteristics, and the change relative to any of the previous periods.

Most of the analysis of the above designs concentrates on efficiency of the design in terms of variance. Another important practical consideration is the implementation cost (Groves 1989). The traditional cost models such as those used in derivation of Neyman-Tchuprow optimal allocation design (Neyman 1938) will now have to include terms related to the cost of the first visit to the cluster and/or ultimate interview unit, as well as the cost of consecutive visits, and the cost of follow-up. It can be argued that the costs of interviewing individuals on the occasions after the first one are greater than on the first occasion (the household may have moved out, so it needs to be located again in another part of the geographic area; they may be unwilling to be interviewed again; etc.), while the cost of revisiting the cluster may be lower (there is no need to do the maps all over again; the cooperation with the community leaders and/or individual respondents has already been established earlier; accuracy of the previously collected data may be improved on revisits).

An additional concern almost inevitably raised in economic literature, especially in the area of program evaluation, is the issue of endogeneity and unobservable characteristics. For that purpose, the panel studies provide an excellent tool to condition on the unobserved characteristics with the fixed effects and difference-in-difference (DID) estimators. A general review of the program evaluation methods can be found in Wooldridge (2002, Ch. 18), and for DID methods in particular, see Bertrand, Duflo & Mullainathan (2004). The papers that were taking into account the cluster structure of the sample and cluster level characteristics include Angeles, Guilkey & Mroz (1998) and Pitt, Rozenzweig & Gibbons (1993).

This paper was motivated by the design of Demographic and Health Surveys[2], 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 particular intertemporal aspect of the design is that the same 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, while the second stage samples are taken independently.

This paper is devoted to the analysis of the designs that feature the repeated use of the clusters from the master sample of PSUs, like in the design of Demographic and Health Surveys. We shall refer to such designs as *cluster-panel*

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

designs. and is structured as follows. Following the introduction (Section 1), Section 2 describes a very basic setting of a simple random sampling (SRS) repeated on two occasions, and Section 3 adds the cost considerations for this design. Then Section 4 introduces clustered designs, Section 5 considers samples taken at two occasions, and Section 6 derives the costs of those designs. Sections 7 and 8 illustrate the considered designs and estimators with an empirical example based on DHS data and on Monte Carlo simulations. Section 9 concludes.

2 Simple setting: one-stage sampling with replacement

The simplest setting to be analyzed is that of two samples from the same population of individuals with some overlap between them admitted by design. Suppose a sample of n individuals is taken *with replacement*[3] from the population of size N at time $t = 1$. Then the mean of the quantity of interest y_1 is estimated as

$$\bar{y}_1 = \frac{1}{n} \sum_{i=1}^{n} y_{1i} \tag{2.1}$$

It is an unbiased estimator of the population quantity

$$\bar{Y}_1 = \frac{1}{N} \sum_{j=1}^{n} y_{1j} \tag{2.2}$$

with variance

$$\mathbb{V}[\bar{y}_1] = \frac{1}{n} S_1^2 \tag{2.3}$$

where

$$S_1^2 = \frac{1}{N-1} \sum_{j=1}^{N} \bigl(y_{1j} - \bar{Y}_1\bigr)^2 \tag{2.4}$$

is the population variance of the quantity.

Suppose the second sample is taken in the next period of time, $t = 2$. Suppose also that there is no change in the population composition, and the characteristics y_{2j} of the population individuals change in such a way that the population variance

$$S_2^2 = \frac{1}{N-1} \sum_{j=1}^{N} \bigl(y_{2j} - \bar{Y}_2\bigr)^2 \tag{2.5}$$

[3] Realistically, most samples are performed without replacement (WOR). The with replacement (WR) assumption, however, simplifies the analytic arguments and provide conservative variance estimates. The difference between the variances obtained from samples with and without replacement can usually be expressed in terms of the sampling fractions $f = n/N$, so for one-stage simple random sample, the relation between the two variances is $\mathbb{V}_{\mathrm{WOR}}[\bar{x}] = (1 - f)\, \mathbb{V}_{\mathrm{WR}}[\bar{x}]$. If the sampling fraction is small (say $f < 10\%$) as in most large surveys, the difference can be ignored.

does not change, $S_2^2 = S_1^2 = S^2$. (A simple case is when $y_{2j} = y_{1j} + a$ for some constant a.)

Suppose we are going to take the sample of the same size, n, in the second period. The design option that the researcher can control is the overlap of the new sample with the sample at time $t = 1$. Let us denote the proportion of observation taken from the first sample as $0 \leq \pi \leq 1$, and rearrange observations 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. [4] Then a natural unbiased estimator of the change $Y_2 - Y_1$ is the difference of the sample averages (also called an *elementary estimate*, c.f. composite estimate later):

$$\hat{\delta} = \bar{y}_2 - \bar{y}_1 \tag{2.6}$$

and it can be decomposed into terms corresponding to independent and overlapping samples:

$$\hat{\delta} = \frac{1}{n}\Big(- \sum_{i=1}^{(1-\pi)n} y_{1i} + \sum_{i=(1-\pi)n+1}^{n} (y_{2i} - y_{1i}) + \sum_{i=n+1}^{(2-\pi)n} y_{2i} \Big) \tag{2.7}$$

By using the fact that the subsamples are independent from each other the variance of this estimator can be obtained as follows:

$$\begin{aligned}\mathbb{V}\big[\hat{\delta}\big] &= \frac{1}{n^2}\,\mathbb{E}\left(- \sum_{i=1}^{(1-\pi)n} (y_{1i} - \bar{Y}_1) + \sum_{i=(1-\pi)n+1}^{n} (y_{2i} - y_{1i} - \bar{Y}_2 + \bar{Y}_1) + \sum_{i=n+1}^{(2-\pi)n} (y_{2i} - \bar{Y}_2) \right)^2 \\ &= \frac{1}{n^2}\,\mathbb{E}\left(\sum_{i=1}^{(1-\pi)n} (y_{1i} - \bar{Y}_1)^2 + \sum_{i=(1-\pi)n+1}^{n} (y_{2i} - y_{1i} - \bar{Y}_2 + \bar{Y}_1)^2 + \sum_{i=n+1}^{(2-\pi)n} (y_{2i} - \bar{Y}_2)^2 \right) = \\ &= \frac{1}{n^2}\big[(1-\pi)nS^2 + \pi n 2S^2(1-\rho) + (1-\pi)nS^2\big] = \frac{2(1-\pi\rho)S^2}{n} \end{aligned} \tag{2.8}$$

where ρ is the correlation of the measurements over time,

$$\frac{1}{N-1}\sum_{j=1}^{N}(y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2) = \rho S^2 \tag{2.9}$$

For any level of ρ, the lowest variance is achieved by setting $\pi = 1$, i.e., by conducting a panel study.

Other possible estimators of the mean and change can be found in the class of *composite estimators* (Hansen, Hurwitz & Madow 1953, Rao & Graham 1964,

[4] The assumption of sampling with replacement is crucial here, as the composition of two independent WR samples is a WR sample. However, the composition of two independent WOR samples is not guaranteed to be a sample without replacement.

Wolter 1979) that use information from several observation occasions to estimate the contemporary mean:

$$\bar{y}_{2\alpha} = (1-\alpha)\bar{y}_2 + \alpha(\bar{y}_{2,\pi} + \bar{y}_{1\alpha} - \bar{y}_{1,\pi}) \tag{2.10}$$

where subindices $1,\pi$ and $2,\pi$ denote the sample means taken over the overlapping parts of the sample. As long as no observations were available prior to $t=1$, $\bar{y}_{1\alpha} = \bar{y}_1$, so the estimator would be better referred to as *one-step composite estimator*. The sampling variance of this estimator can be found as

$$\begin{aligned}
\mathbb{V}\big[\bar{y}_{2\alpha}\big] = \mathbb{V}\Big[\frac{1-\alpha}{n}\sum_{i=(1-\pi)n+1}^{(2-\pi)n} y_{2i} + \frac{\alpha}{\pi n}\sum_{i=(1-\pi)n+1}^{n} y_{2i} + \frac{\alpha}{n}\sum_{i=1}^{n} y_{1i} - \frac{\alpha}{\pi n}\sum_{i=(1-\pi)n+1}^{n} y_{1i}\Big] = \\
= \mathbb{V}\Big[\sum_{i=n}^{(2-\pi)n}\frac{1-\alpha}{n}y_{2i} + \sum_{i=(1-\pi)n+1}^{n}\Big(\frac{\alpha}{\pi n}+\frac{1-\alpha}{n}\Big)y_{2i} + \\
+ \sum_{i=1}^{(1-\pi)n}\frac{\alpha}{n}y_{1i} - \sum_{i=(1-\pi)n+1}^{n}\frac{\alpha(1-\pi)}{\pi n}y_{1i}\Big] = \\
= \frac{S^2}{n}\big[(1-\alpha)^2(1-\pi) + \alpha^2(1-\pi) + \\
+\pi^{-1}\big((\alpha+\pi-\alpha\pi)^2 + \alpha^2(1-\pi)^2 - 2\rho(\alpha+\pi-\alpha\pi)\alpha(1-\pi)\big)\big]
\end{aligned} \tag{2.11}$$

It can be minimized over α to achieve the optimal design:

$$\begin{aligned}
\frac{\partial\,\mathbb{V}\big[\bar{y}_{2\alpha}\big]}{\partial\alpha} = \frac{2S^2}{n}\big[-(1-\alpha)(1-\pi) + \alpha(1-\pi) + \\
+\pi^{-1}\big((1-\pi)(\alpha+\pi-\alpha\pi) + \alpha(1-\pi)^2 - \rho(1-\pi)(2\alpha+\pi-2\alpha\pi)\big)\big], \\
\alpha^* = \frac{\rho\pi}{2(1-\rho(1-\pi))}
\end{aligned} \tag{2.12}$$

This can be plugged back to (2.10) and (2.11) for the optimal estimate of the mean and its variance. Jessen (1942) found that the optimal overlap between two samples is achieved at approximately

$$\frac{\pi}{1-\pi} = (1-\rho^2)^{1/2} \tag{2.13}$$

and reported efficiency gains between 22% and 45% in his agricultural application. Note that when $\rho = 0$ or $\pi = 0$, the estimator boils down to the usual mean of the second sample.

A one-step composite estimate of the change is the following:

$$\hat{\delta}_\alpha = \frac{1}{n}\Big[(1+\pi\alpha)\Big(-\sum_{i=1}^{(1-\pi)n} y_{1i} + \sum_{i=n+1}^{(2-\pi)n} y_{2i}\Big) + (1-(1-\pi)\alpha)\sum_{i=(1-\pi)n+1}^{n}(y_{2i}-y_{1i})\Big] \tag{2.14}$$

for some α. When $\alpha = 0$, the estimator coincides with the elementary estimate (2.6). The weights $1+\pi\alpha$ and $1-(1-\pi)\alpha$ ensure that the estimator is unbiased for the change in population means, and the variance can be derived, similarly to (2.8), as

$$\mathbb{V}\big[\hat{\delta}_\alpha\big] = \frac{2S^2}{n}\big[(1+\pi\alpha)^2(1-\pi) + (1-(1-\pi)\alpha)^2\pi(1-\rho)\big] \tag{2.15}$$

This expression can now be minimized over π and α:

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

$$\alpha^*(\rho,\pi) = -\frac{\rho}{1-\rho(1-\pi)} \tag{2.16}$$

When this optimal α^* is substituted into (2.15), the variance of interest becomes

$$\mathbb{V}\big[\hat{\delta}_{\alpha^*}\big] = \frac{2S^2}{n}\frac{1-\rho}{1-\rho(1-\pi)} \tag{2.17}$$

which is decreasing in π over $[0,1]$. Thus the minimum is again attained for the panel setting with all observations retained from the first round of observation. Then the first two sums in (2.14) disappear, and the one-step composite estimator coincides with the elementary estimate (2.6).

For $\pi < 1$, the one-step composite estimator is more efficient than (2.6); the difference is of the order $O(\rho^2)$ for small ρ. The efficiency gain is greater than 20% only for relatively high correlations ($\rho > 0.63$); in the extreme case of $\rho = 1$, the one-step composite estimator has a variance of zero, while the elementary estimate still has non-zero variance.

The efficiency advantages of the rotation designs and corresponding composite estimators are, at least to some extent, counterweighed by the *rotation bias*, or *conditioning*, which is the effect of presence in the sample for several occasions that changes the behavior of the observed unit. Thus the estimates coming from different panels of a rotation sample may not have the same statistical properties. This bias is a part of non-sampling error and thus is outside of the randomization based inference paradigm. The issue may be less of a problem with DHS as it only reuses the clusters rather than individual observations in the sample, and the period of time between consecutive rounds of data collection is quite large (5 years).

3 One-stage sampling with replacement and varying costs

The preceding analysis does not take into account the possibility of varying costs in two periods of time. 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:

$$\frac{2(1-\pi\rho)S^2}{n} \to \min_{n,\pi}$$
$$\text{s.t. } c_1(1-\pi)n + c_{12}\pi n + c_2(1-\pi)n \leq C_0 \tag{3.1}$$

where C_0 is the part of the budget allowed for the varying costs. For a given π, the sample size can be found from the cost equation:

$$n = \frac{C_0}{c_1 + c_2 + \pi(c_{12} - c_1 - c_2)} \tag{3.2}$$

so the minimization equation becomes

$$V(\pi;\rho) = \frac{2(1-\pi\rho)(c_1 + c_2 + \pi[c_{12} - c_1 - c_2])}{C_0} \to \min_{\pi\in[0,1]} \tag{3.3}$$

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

$$\pi^* = \frac{c_{12} - c_1(1+\rho) - c_2(1+\rho)}{2\rho(c_{12} - c_1 - c_2)} = \frac{1}{2\rho} - \frac{1}{2\left[\frac{c_{12}}{c_1+c_2} - 1\right]} \tag{3.4}$$

Let us now consider three special cases.

1. $c_{12} < c_1 + c_2$. Even though this relationship does not seem plausible, since tracking individuals tends to add extra expenses to the survey budget, it will make more sense for the clustered samples when it might be cheaper to revisit the already sampled location than to prepare the maps and the frame for a new one.
   In this case, π^* gives the location of the minimum. As long as $\rho \leq 1$, and $0 < c_{12} < c_1 + c_2$, both of the terms in the last equality of (3.4) are greater than $1/2$, so $\pi^* > 1$, and the objective function is decreasing to the left of 1. Hence the minimum over $[0, 1]$ is achieved at $\pi = 1$.

2. $c_{12} = c_1 + c_2$, there are neither extra costs nor extra savings associated with the panel mode of data collection. In this case the objective function is linear: $\mathbb{V}[\hat{\delta}] = 2(1-\pi\rho)(c_1 + c_2)/C_0$, and the optimal solution to it is $\pi = 1$. This case is equivalent to the one considered in the previous section where the question of costs did not arise.

3. $c_{12} > c_1 + c_2$. For this most realistic case, when the logistic of finding the same individuals for a panel study implies additional costs, the objective function in (3.3) is a parabola with downward branches, and π^* given by (3.4) determines the maximum of it. Hence the minimization problem (3.3) has a corner solution at either 1 or 0, depending on which of those two points is further from the maximum.

A simple way to solve the problem is then to compare those two corner solutions:

$$V(0;\rho) = \frac{2(c_1+c_2)}{C_0},$$
$$V(1;\rho) = \frac{2(1-\rho)c_{12}}{C_0} \tag{3.5}$$

so the optimal solution is then

$$\pi = \begin{cases} 0, & c_1+c_2 < (1-\rho)c_{12}, \text{ or } \rho < 1 - \dfrac{c_1+c_2}{c_{12}} \\ 1, & c_1+c_2 > (1-\rho)c_{12}, \text{ or } \rho > 1 - \dfrac{c_1+c_2}{c_{12}} \end{cases} \tag{3.6}$$

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

The last case shows that the sample designer must have an accurate understanding of the underlying data process to decide between collecting the data from two completely independent samples ($\pi = 0$, small ρ) and collecting the panel data ($\pi = 1$, high ρ).

For the one-step composite estimator of the difference (2.14), the minimization problem is

$$\frac{2S^2}{n}\frac{1-\rho}{1-\rho(1-\pi)} \to \min_{n,\pi}$$
$$\text{s.t. } c_1(1-\pi)n + c_{12}\pi n + c_2(1-\pi)n \le C_0 \tag{3.7}$$

Making the same substitution for n, one gets the equivalent minimization problem

$$\frac{2S^2}{C_0}\frac{(1-\rho)(c_1+c_2+\pi(c_{12}-c_1-c_2))}{1-\rho(1-\pi)} \to \min_{\pi} \tag{3.8}$$

The first derivative of the objective function is

$$\frac{2S^2}{C_0}\frac{(1-\rho)((1-\rho)c_{12}-c_1-c_2)}{(1-\rho(1-\pi))^2} \tag{3.9}$$

which is negative for $(1-\rho)c_{12} < c_1+c_2$, and hence the panel design is optimal; and positive for $(1-\rho)c_{12} > c_1+c_2$, and hence independent sampling is optimal. The elementary and one-step composite estimates thus have the same design efficiency requirements, as the above conditions are identical to those considered in (3.6) for the elementary estimate.

4 Clustered samples

In this section, we describe the basic aspects of variance and its estimation for a two-stage clustered sample.

A sample is said to be *clustered* if the primary sampling unit (PSU) is not the observation unit, but a group of such units. Samples are usually collected in the clustered manner if no listing of individual units is available, or such listing is prohibitively difficult or expensive to obtain. Instead, geographic areas with a large number of observation units are enumerated and sampled, possibly with unequal selection probabilities. Further, all units of a cluster may be observed, or samples can be taken within those clusters, either in the cluster form again, or as a SRS if a list of units can finally be obtained. In DHS studies, the clusters (PSUs) are the communities, the secondary sampling units (SSUs) are the dwelling units, and all individuals are observed within the SSUs.

Why are the clustered samples used in practice? Typically, the variances obtained from clustered samples are higher than from an SRS of the same size, because units within clusters tend to be similar, and thus measurements taken on those units tend to be positively correlated. Since there is not enough variability among them, the effective number of distinct observations is smaller than the sample size. The measure of differences is referred to as the *design effect* (Kish 1965):

$$DEFF = \frac{\mathbb{V}[\text{design}, n]}{\mathbb{V}[\text{SRS}, n]} \tag{4.1}$$

It can be shown (see Thompson 1992, p. 122; Hansen, Hurwitz and Madow, 1953, vol. II, sec. 6.5; or Kish (1965)) that the design effect for the clustered sample is approximately

$$DEFF \approx 1 + \rho(\bar{m} - 1) \tag{4.2}$$

where $\bar{m}$ is the average number of units per cluster, and ρ is the measure of cluster homogeneity, or within-cluster correlation, or *intraclass correlation.* See the operational definition below in (4.15).

The primary advantage of the clustered samples, however, is that they are usually substantially cheaper to administer than SRS, yet still give reasonable precision if designed with the above considerations on design effects are kept in mind. A successful clustered sample will have design effects smaller than 2, meaning that the increase of the variance relative to the simple random sample of the same size is at most by a factor of two, or that the researchers had to take at most twice as many observations to achieve a desirable margin of error. Unfortunately, for community-level variables, such as access to electricity, the design effects will inevitably be above 10 reflecting the fact that the individuals within the community are almost perfectly correlated (typically, nearly everybody has electric power, or practically no one has), and the effective sample size is then the number of PSUs.

For the following analysis, assume a two-stage cluster equal probability of selection (epsem) design where the finite population corrections may be ignored (i.e., either the sampling is with replacement, or sampling fractions at each stage are small. The resulting variance formulae will be conservative for real samples without replacement). Such design can be obtained by selecting SRS at both stages; more complicated probability designs will produce conceptually the same results attenuated by weights.

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 $j = 1, \ldots, M_i$ for i-th cluster in population, and $j = 1, \ldots, m_i$ in the sample. The sampling fractions are $f_i = m_i/M_i$ within clusters, and $f_I = n/N$ is the fraction of clusters sampled. 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 can then be found as follows:

$$T[Y]_{i\cdot} = Y_{i\cdot} = \sum_{j=1}^{M_i} Y_{ij} \,, \tag{4.3}$$

$$t[y]_{i\cdot} = y_{i\cdot} = \frac{M_i}{m} \sum_{j=1}^{M_i} y_{ij} \,, \tag{4.4}$$

$$T[Y]_{\cdot\cdot} = \sum_{i=1}^{N} Y_{i\cdot} \,, \tag{4.5}$$

$$t[y]_{\cdot\cdot} = \frac{N}{n} \sum_{i=1}^{N} y_{i\cdot} \tag{4.6}$$

Y_{ij} can be any individual level characteristic. One special case is $Y_{ij} = 1$, and $T[1]_{\cdot\cdot}$ is the total population size, $T[1]_{i\cdot} = M_i$ is the size of i-th cluster, and corresponding t are their (unbiased) estimators.

The means per observation units are

$$\bar{Y}_{i\cdot} = \frac{1}{M_i} \sum_{i=1}^{M_i} Y_{ij} = \frac{Y_{i\cdot}}{M_i}, \tag{4.7}$$

$$\bar{y}_{i\cdot} = \frac{1}{M_i} \sum_{i=1}^{M_i} y_{ij} = \frac{Y_{\cdot\cdot}}{N\bar{M}} \tag{4.8}$$

$$\bar{Y}_{\cdot\cdot} = \frac{T[Y]_{\cdot\cdot}}{N\bar{M}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}}{N\bar{M}} \tag{4.9}$$

$$\bar{y}_{\cdot\cdot} = \frac{t[y]_{\cdot\cdot}}{n\bar{m}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}}{n\bar{m}} \tag{4.10}$$

where

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i \tag{4.11}$$

is the average size of the cluster, with the appropriate sample analogue $\bar{m}$.

Note that the estimator of the mean (4.10) is a *ratio* estimator: both the numerator and the denominator are random variables design-unbiased for their population analogues.[5] As it is a nonlinear function of unbiased estimators, it is no longer unbiased for the population quantity it estimates, but it is a consistent estimator as the number of clusters $n \to \infty$, and is approximately unbiased in large samples.

The variance of Y and its within- and between-cluster components are

$$S^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2}{N\bar{M} - 1}, \tag{4.12}$$

$$S^2_{wi} = \frac{\sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_{i\cdot})^2}{M_i - 1}, \tag{4.13}$$

$$S^2_b = \frac{\sum_{i=1}^{N} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{N - 1} \tag{4.14}$$

The *intraclass correlation*, one of the determinants of the design effect (4.2), is the share of variance due to clustering:

$$\rho = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{j \neq j'} (Y_{ij} - \bar{Y}_{\cdot\cdot})(Y_{ij'} - \bar{Y}_{\cdot\cdot})}{(\bar{M} - 1)(N\bar{M} - 1)S^2} = \frac{(N-1)\bar{M}^2 S^2_b - N(\bar{M} - 1)S^2_w}{(\bar{M} - 1)(N\bar{M} - 1)S^2} \approx$$

$$\approx \frac{S^2_b - S^2/\bar{M}}{(\bar{M} - 1)S^2/\bar{M}} \tag{4.15}$$

when $N \gg 1$.

Variance of the total $t_{\cdot\cdot}$ is (Hansen et. al. 1953, v.2, sec. 6.1; Särndal et. al., 1992, sec. 4.3)

$$\mathbb{V}[t_{\cdot\cdot}] = N^2 \frac{1 - f_I}{n} S^2_b + N/n \sum_{i=1}^{n} M^2_i \frac{1 - f_i}{m_i} S^2_{wi} \tag{4.16}$$

The most traditional analytical way to obtain the variance (or, to be precise, the mean squared error) of a ratio estimator such as (4.10) is to use the linearization, also known as Taylor series expansion, or the delta method. If a ratio $R = T[Y]/T[X]$ of two totals is estimated by $r = t[y]/t[x]$, then the first order expansion is given by

[5] It is rather hard to think of the sample size as a random variable, but it is such in most survey contexts if a fixed number of observations per cluster is not specified in advance. Variability in the sample size is a subtle but important component of the overall variance.

$$r(t[x],t[y]) = t[y]/t[x] = R(T[Y],T[X]) + \frac{\partial R}{\partial T[X]}(t[x]-T[X]) + \frac{\partial R}{\partial T[Y]}(t[y]-T[Y]) + o_p(t[x]-T[X],t[y]-T[Y]) \qquad (4.17)$$

so upon computing the appropriate derivatives

$$\frac{\partial R}{\partial T[X]} = -\frac{T[Y]}{T[X]^2}, \quad \frac{\partial R}{\partial T[Y]} = \frac{1}{T[X]} \qquad (4.18)$$

the variance (MSE) of the estimate can be found as

$$\mathrm{MSE}[r] = \mathbb{E}[r-R]^2 \approx \left(\frac{T[Y]}{T[X]}\right)^2 \mathbb{V}\big[t[x]\big] + \frac{1}{T[X]^2}\,\mathbb{V}\big[t[y]\big] - 2\frac{T[Y]}{T[X]^3}\,\mathbb{C}\mathrm{ov}\big[t[y],t[x]\big] \qquad (4.19)$$

and an estimate of this variance is found by replacing the totals, variances and covariances by their unbiased estimators. In estimation of the mean per unit, the second variable, Y, is set to 1, so its variance is the variance of the estimator of the total population size,

$$\mathbb{V}\big[t[1]\big] = \frac{1}{N-1}\sum_{i=1}^{N} \frac{M_i^2}{m_i^2}(m_i - \bar{m})^2 \qquad (4.20)$$

This shows an important fact in cluster sampling: the number of units per cluster should be about the same to minimize this component of variance. A sampling scheme that controls the average cluster size should result in low overall variance.

One such scheme is *probability proportional to size* method of selection (PPS). If the sizes of the clusters M_i are known before sampling, or some measure of cluster size is available instead, the probability of selecting i-th cluster can be taken as

$$\mathrm{Prob}[i] = \frac{n}{N}M_i \qquad (4.21)$$

and probability of selecting a unit j from selected cluster i, as

$$\mathrm{Prob}[j|i] = m/M_i \qquad (4.22)$$

where m is the fixed number of units selected from each cluster. Then the total probability of selection for unit j in i-th cluster is

$$\mathrm{Prob}[ij] = \frac{n}{N}M_i\frac{m}{M_i} = \frac{nm}{N} = \mathrm{const} \qquad (4.23)$$

is the same for all ultimate sampling units, and the design is both self-weighting (epsem), and efficient for variance, as it eliminates the component of variance related to the variability in the sample size.

If we make an additional assumption that the sampling fractions are the same across clusters ($f_i = f_{II}$), the variance of the mean $y_{\cdot\cdot}$ can be obtained as the variance of the ratio estimator by linearization (or delta) method; see Korn & Graubard (1999, p. 27), Thompson (1992, p. 116):

$$\mathbb{V}[y_{\cdot\cdot}] \approx \frac{1}{\bar{m}^2 N(n-1)} \sum_{i=1}^{N} m_i^2 (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \tag{4.24}$$

If we compare it to the naïve estimate implied by the simple random sample,

$$\mathbb{V}[\text{SRS}, n] = \frac{S^2}{n} \tag{4.25}$$

the design effect can be obtained in the form (4.2). Note that if clusters are similar to one another (S_b^2 is small), and there is a lot of variability within each cluster (S_{wi}^2 are all large), then $\rho < 0$, and clustered sampling design is in fact more efficient than SRS. In that case, the optimal design is to sample as few clusters as possible, and take as many observations per cluster as possible. In most practical cases, however, the clusters can be expected to be rather homogeneous (in terms of population sampling, think of differences within the community vs. differences between communities), so a better design would be to sample many clusters and a few observations per cluster.

The relation between the cluster size and the resulting design effect is somewhat more complicated than may seem from (4.2) as increase in the cluster size usually leads to a greater diversity of the units within it, and hence to smaller within-cluster correlations. Thompson (1992, Sec. 12.5) discusses the size of the intraclass correlation that appears in (4.15), and argues that ρ is higher in small clusters, and in relatively compact (say square) clusters compared to needle-like clusters that have units further apart, and thus less correlated.

5 Repeated cluster samples

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: $Y_{ij}^{(t)}$, $t = 1, 2, \ldots$

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

$$D[\bar{Y}^{(2)} - \bar{Y}^{(1)}] = \frac{T[Y^{(2)}]}{T[1^{(2)}]} - \frac{T[Y^{(1)}]}{T[1^{(1)}]} \tag{5.1}$$

This is often an important policy measure, such as changes in the contraceptive use, mortality or fertility between years of data collection. The time notation on the unit variable shows that the population size may change over time.

In terms of the previous section, expression (5.1) 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.

For general ratio estimators of y/x, the estimator of (5.1), although biased in finite samples, is the difference of the corresponding ratio estimators:

$$d[\bar{y}^{(2)} - \bar{y}^{(1)}] = \frac{t[y^{(2)}]}{t[x^{(2)}]} - \frac{t[y^{(1)}]}{t[x^{(1)}]} = d\bigl(t[y^{(2)}], t[y^{(1)}], t[x^{(2)}], t[x^{(1)}]\bigr) \tag{5.2}$$

and its variance is

$$\mathbb{V}\bigl[d[y^{(2)} - y^{(1)}]\bigr] = \mathbb{V}\Bigl[\frac{T[Y^{(2)}]}{T[X^{(2)}]}\Bigr] + \mathbb{V}\Bigl[\frac{T[Y^{(1)}]}{T[X^{(1)}]}\Bigr] - 2\,\mathbb{C}\mathrm{ov}\Bigl[\frac{T[Y^{(2)}]}{T[X^{(2)}]}, \frac{T[Y^{(1)}]}{T[X^{(1)}]}\Bigr] \tag{5.3}$$

The linear approximations for the first two terms of (5.3) are given by (4.19). 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.

Rather than trying to compute the covariance term explicitly, we can take the first order Taylor series expansion:

$$\begin{gathered}\mathbb{V}\bigl[d[y^{(2)} - y^{(1)}]\bigr] \approx \nabla^T d\bigl(T[Y^{(2)}], T[Y^{(1)}], T[X^{(2)}], T[X^{(1)}]\bigr) \times \\ \times\,\mathbb{C}\mathrm{ov}\Bigl[\bigl(T[Y^{(2)}], T[Y^{(1)}], T[X^{(2)}], T[X^{(1)}]\bigr)^T\Bigr] \nabla d\bigl(T[Y^{(2)}], T[Y^{(1)}], T[X^{(2)}], T[X^{(1)}]\bigr) \equiv \\ \equiv D^T C D\end{gathered} \tag{5.4}$$

where $\nabla d\bigl(t[y^{(2)}], t[y^{(1)}], t[x^{(2)}], t[x^{(1)}]\bigr) \equiv D$ is the gradient of $d(\cdot)$ with entries

$$\begin{gathered}\frac{\partial d}{\partial t[y^{(2)}]} = \frac{1}{t[x^{(2)}]}, \quad \frac{\partial d}{\partial t[y^{(1)}]} = -\frac{1}{t[x^{(1)}]}, \\ \frac{\partial d}{\partial t[x^{(2)}]} = -\frac{t[y^{(2)}]}{t[x^{(2)2}]}, \quad \frac{\partial d}{\partial t[x^{(1)}]} = \frac{t[y^{(1)}]}{t[x^{(1)2}]}\end{gathered} \tag{5.5}$$

and

$$C = \mathbb{C}\mathrm{ov}\Bigl[\bigl(t[y^{(2)}], t[y^{(1)}], t[x^{(2)}], t[x^{(1)}]\bigr)^T\Bigr] \tag{5.6}$$

is the covariance matrix of the four variate vector of totals. The approximation in (5.4) is due to linearization, and all terms in it are evaluated at the population totals.

In particular, the cross-covariance term of (5.3) can now be found as

$$\mathbb{C}\mathrm{ov}\Big[\frac{t[y^{(2)}]}{t[x^{(2)}]},\frac{t[y^{(1)}]}{t[x^{(1)}]}\Big] \approx$$

$$\approx \frac{\partial d}{\partial t[y^{(2)}]}\frac{\partial d}{\partial t[y^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[y^{(2)}],t[y^{(1)}]\big] + \frac{\partial d}{\partial t[y^{(2)}]}\frac{\partial d}{\partial t[x^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[y^{(2)}],t[x^{(1)}]\big]$$

$$+\frac{\partial d}{\partial t[x^{(2)}]}\frac{\partial d}{\partial t[y^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[x^{(2)}],t[y^{(1)}]\big] + \frac{\partial d}{\partial t[x^{(2)}]}\frac{\partial d}{\partial t[x^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[x^{(2)}],t[x^{(1)}]\big] =$$

$$= \frac{1}{T[X^{(1)}]T[X^{(2)}]}\Bigg\{\mathbb{C}\mathrm{ov}\big[t[y^{(2)}],t[y^{(1)}]\big] - \frac{T[Y^{(1)}]}{T[X^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[y^{(2)}],t[x^{(1)}]\big]$$

$$-\frac{T[Y^{(2)}]}{T[X^{(2)}]}\,\mathbb{C}\mathrm{ov}\big[t[x^{(2)}],t[y^{(1)}]\big] + \frac{T[Y^{(2)}]}{T[X^{(2)}]}\frac{t[y^{(1)}]}{t[x^{(1)}]}\,\mathbb{C}\mathrm{ov}\big[t[x^{(2)}],t[x^{(1)}]\big]\Bigg\} \qquad (5.7)$$

By the law of iterated expectations, each of covariances in the last expression is of the form

$$\mathbb{C}\mathrm{ov}\big[t[\xi^{(2)}],t[\zeta^{(1)}]\big] =$$

$$= \mathbb{E}_{\mathrm{I}}\Big\{\mathbb{C}\mathrm{ov}_{\mathrm{II}}\big[t[\xi^{(2)}],t[\zeta^{(1)}]|\mathrm{I}\big]\Big\} + \mathbb{C}\mathrm{ov}_{\mathrm{I}}\Big\{\mathbb{E}_{\mathrm{II}}\big[t[\xi^{(2)}]|\mathrm{I}\big],\mathbb{E}_{\mathrm{II}}\big[t[\zeta^{(1)}]|\mathrm{I}\big]\Big\} \qquad (5.8)$$

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

$$\mathbb{E}_{\mathrm{II}}\big[t[\xi^{(2)}]|\mathrm{I}\big] = \frac{N}{n}\sum_{i=1}^{n}\Xi_{i\cdot}^{(2)} \qquad (5.9)$$

is the estimate of the total based on the population mean from clusters in the sample. (There is a subtle abuse of notation here: the capitalized quantities refer to the population parameters, but here we have to mix together the population parameters (cluster totals) $\Xi_{i\cdot}$ and the sample elements indexed by i, due to the conditioning on the particular sample of PSUs chosen at the first stage.) Suppose a fraction of π of the original clusters are reused in the second wave of the survey (c.f. Section 2). Then

$$\mathbb{C}\mathrm{ov}\big[t[\xi^{(2)}],t[\zeta^{(1)}]\big] = \mathbb{C}\mathrm{ov}_{\mathrm{I}}\Big\{\mathbb{E}_{\mathrm{II}}\big[t[\xi^{(2)}]|\mathrm{I}\big],\mathbb{E}_{\mathrm{II}}\big[t[\zeta^{(1)}]|\mathrm{I}\big]\Big\} =$$

$$= \frac{N^2\pi n}{n^2}\,\mathbb{C}\mathrm{ov}\big[\Xi_{i\cdot}^{(2)},\mathrm{Z}_{i\cdot}^{(1)}\big] = \frac{\pi N^2}{(N-1)n}\sum_{i=1}^{N}(\Xi_{i\cdot}^{(2)}-\bar{\Xi}_{i\cdot}^{(2)})(\mathrm{Z}_{i\cdot}^{(1)}-\bar{\mathrm{Z}}_{i\cdot}^{(1)}) \qquad (5.10)$$

is the population covariance between the cluster means over two time periods, where

$$\bar{\Xi}_{i\cdot}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\Xi_{i\cdot}^{(t)} \approx \bar{M}\bar{\Xi}_{\cdot\cdot}^{(t)} \qquad (5.11)$$

etc., and the latter approximate equality holds if clusters are about the same size $\bar{M}$.

For further analysis, we shall return to the case when $x_{ij} = 1$, and also make a simplifying assumption that the design is fixed size, so $\mathbb{V}\bigl[t[x^{(t)}]\bigr] = 0$, $t = 1, 2, \ldots$. Then the covariance term in (5.3) is comprised only of the covariances between y's in two time periods:

$$\mathbb{C}\mathrm{ov}\Bigl[\frac{t[y^{(2)}]}{t[x^{(2)}]}, \frac{t[y^{(1)}]}{t[x^{(1)}]}\Bigr] \approx \frac{\mathbb{C}\mathrm{ov}\bigl[t[y^{(2)}], t[y^{(1)}]\bigr]}{T[X^{(1)}]T[X^{(2)}]} \approx$$

$$\approx \frac{N^2 \pi\, \mathbb{C}\mathrm{ov}\bigl[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\bigr]}{(N\bar{M})^2 n} = \frac{\pi\, \mathbb{C}\mathrm{ov}\bigl[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\bigr]}{\bar{M}^2 n} \tag{5.12}$$

where the first approximation is the linearization / Taylor series expansion for the covariance in question, and the second one is the linearization / Taylor series expansion for the covariance in the numerator.

Combining this result with (5.3), we obtain

$$\mathbb{V}\bigl[d[y^{(2)} - y^{(1)}]\bigr] =$$

$$= \frac{1}{T[X^{(2)}]^2}\, \mathbb{V}\bigl[t[y^{(2)}]\bigr] + \left(\frac{T[Y^{(2)}]}{T[X^{(2)}]}\right)^2 \mathbb{V}\bigl[t[x^{(2)}]\bigr] - 2\frac{T[Y^{(2)}]}{T[X^{(2)}]^3}\, \mathbb{C}\mathrm{ov}\bigl[t[x^{(2)}], t[y^{(2)}]\bigr] +$$

$$+ \frac{1}{T[X^{(1)}]^2}\, \mathbb{V}\bigl[t[y^{(1)}]\bigr] + \left(\frac{T[Y^{(1)}]}{T[X^{(1)}]}\right)^2 \mathbb{V}\bigl[t[x^{(1)}]\bigr] - 2\frac{T[Y^{(1)}]}{T[X^{(1)}]^3}\, \mathbb{C}\mathrm{ov}\bigl[t[x^{(1)}], t[y^{(1)}]\bigr] -$$

$$-2\frac{\pi\, \mathbb{C}\mathrm{ov}\bigl[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\bigr]}{\bar{M}^2 n} \tag{5.13}$$

For a fixed size design, $\mathbb{V}[t(x^{(1)})] = \mathbb{V}[t(x^{(1)})] = 0$, and (5.13) simplifies to

$$\mathbb{V}\bigl[d[y^{(2)} - y^{(1)}]\bigr] =$$

$$= \frac{1}{N^2 \bar{M}^2}\, \mathbb{V}\bigl[t[y^{(2)}]\bigr] + \frac{1}{N^2 \bar{M}^2}\, \mathbb{V}\bigl[t[y^{(1)}]\bigr] - 2\frac{\pi\, \mathbb{C}\mathrm{ov}\bigl[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\bigr]}{\bar{M}^2 n} \tag{5.14}$$

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 (5.14), provided the first two terms of (5.14) do not change across the designs with varying π. 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

$$DEFF_r = \frac{\mathbb{V}[\text{repeated design}]}{\mathbb{V}[\text{independent sampling}]} = 1 - 2\frac{\pi\, \mathbb{C}\mathrm{ov}\bigl[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\bigr]}{n(\mathbb{V}\bigl[t[y^{(1)}]\bigr] + \mathbb{V}\bigl[t[y^{(2)}]\bigr])/N^2} \tag{5.15}$$

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 $n \to \infty$.

6 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.

In DHS, at least in the rural areas, clusters arise naturally as villages or communities.

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 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 $t = 2$.

We could have also distinguished the second stage costs for the second period for the individuals sampled from the new or from the reused clusters, but as long as sampling and data collection are performed independently in two time periods, there should not be any major differences in costs (and, for that matter, not a major difference in costs c_1^{II} and c_2^{II}, except for inflation). Also, the costs may depend on the cluster size M_i, as it may take more time and resources to obtain maps and collect cluster level data for bigger clusters.

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

$$C_0 = c_1^{\mathrm{I}} n_1 + c_{12}^{\mathrm{I}} n_0 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}} (n_1 + n_0) m_1 + c_2^{\mathrm{II}} (n_2 + n_0) m_2 \tag{6.1}$$

The sample designer wishes to minimize the variance of the difference of ratio estimators (5.3):

$$\{n_0, n_1, n_2, m_1, m_2\} = \arg\min \mathbb{V}\Big[\frac{t[y^{(2)}]}{t[1^{(2)}]} - \frac{t[y^{(1)}]}{t[1^{(1)}]}\Big] \tag{6.2}$$

Note that this is a rather artificial objective function that focuses solely on the difference between the two sample means. A more realistic design would be finding a balance between minimizing the contemporaneous and longitudinal variances. This would make the optimization problem multicriterial and thus substantially more difficult to handle. Note however that for the contemporaneous estimation, the cluster panel aspect would not matter: the variance of $t[y^{(t)}]/t[1^{(t)}]$, $t = 1, 2$, will be the same for designs with different degree of overlap, and hence the sample designer will choose the relation between n_0, n_1 and n_2 based solely on the cluster panel aspect of variance of $t[y^{(2)}]/t[1^{(2)}] - t[y^{(1)}]/t[1^{(1)}]$. The situation may be different for composite estimators.

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 in (6.2) is

$$\begin{aligned}
\mathbb{V}\Big[\frac{t[y^{(2)}]}{t[1^{(2)}]} - \frac{t[y^{(1)}]}{t[1^{(1)}]}\Big] &= \frac{\mathbb{V}[t[y^{(2)}]}{T[1^{(2)}]^2} + \frac{\mathbb{V}[t[y^{(1)}]}{T[1^{(1)}]^2} - 2\frac{\mathbb{C}\mathrm{ov}\big[t[y^{(2)}], t[y^{(1)}]\big]}{T[1^{(1)}]T[1^{(2)}]} = \\
&= \frac{N - (n_1 + n_0)}{(n_1 + n_0)NM^2} S_{1b}^2 + \frac{1}{NM}\frac{M - m_1}{m_1}\bar{S}_{1w}^2 + \\
+ \frac{N - (n_2 + n_0)}{(n_2 + n_0)NM^2} S_{2b}^2 &+ \frac{1}{NM}\frac{M - m_2}{m_2}\bar{S}_{2w}^2 - 2\frac{n_0 \rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1 + n_0)(n_2 + n_0)M^2}
\end{aligned} \tag{6.3}$$

where the variance (4.16) was used for the first two terms, and

$$\rho^{\mathrm{I}} = \frac{\mathbb{C}\mathrm{ov}\big[Y_{i\cdot}^{(2)}, Y_{i\cdot}^{(1)}\big]}{M^2 S_{1b} S_{2b}} = \frac{1}{(N-1)M^2 S_{1b} S_{2b}} \sum_{i=1}^{N} \big(Y_{i\cdot}^{(2)} - \bar{Y}_{i\cdot}^{(2)}\big)\big(Y_{i\cdot}^{(1)} - \bar{Y}_{i\cdot}^{(1)}\big) \tag{6.4}$$

is the intertemporal correlation of the cluster totals. (Note that the covariance term in the numerator is of order $O(M^2)$, as it is a sum of the cluster totals, hence the scaling factor of M^2 is in place in the denominator.) Also,

$$\bar{S}_{tw}^2 = \frac{1}{N} \sum_{i=1}^{N} S_{twi}^2, \quad t = 1, 2, \dots \tag{6.5}$$

The minimization constraints are:

$$c_1^{\mathrm{I}} n_1 + c_{12}^{\mathrm{I}} n_0 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}}(n_1 + n_0)m_1 + c_2^{\mathrm{II}}(n_2 + n_0)m_2 \le C_0, \tag{6.6}$$

$$n_0 \ge 0, \quad n_1 \ge 0, \quad n_2 \ge 0 \tag{6.7}$$

and with the corresponding Lagrange multipliers λ, λ_0, λ_1, λ_2, the Lagrangian function can be written down as a combination of (6.3), (6.6) and (6.7):

$$\begin{gathered}
L(n_0, n_1, n_2, m_1, m_2; \lambda, \lambda_0, \lambda_1, \lambda_2) = \\
= \Big(\frac{N}{n_1 + n_0} - 1\Big)\frac{S_{1b}^2}{NM^2} + \frac{1}{NM}\Big(\frac{M}{m_1} - 1\Big)\bar{S}_{1w}^2 + \Big(\frac{N}{n_2 + n_0} - 1\Big)\frac{S_{2b}^2}{NM^2} + \\
+ \frac{1}{NM}\Big(\frac{M}{m_2} - 1\Big)\bar{S}_{2w}^2 - \frac{2n_0\rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1 + n_0)(n_2 + n_0)} + \lambda_0 n_0 + \lambda_1 n_1 + \lambda_2 n_2 - \\
- \lambda\big[c_1^{\mathrm{I}} n_1 + c_{12}^{\mathrm{I}} n_0 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}}(n_1 + n_0)m_1 + c_2^{\mathrm{II}}(n_2 + n_0)m_2 - C_0\big] \to \min
\end{gathered} \tag{6.8}$$

The necessary conditions (Karush-Kuhn-Tucker conditions) for the constrained minimum of (6.8) are (Di Pillo & Palagi 2002, Jahn 1996)

$$\frac{\partial L}{\partial n_0} = 0, \quad \frac{\partial L}{\partial n_1} = 0, \quad \frac{\partial L}{\partial n_2} = 0, \tag{6.9}$$

$$\frac{\partial L}{\partial m_1} = 0, \quad \frac{\partial L}{\partial m_2} = 0, \tag{6.10}$$

$$\lambda\big[c_1^{\mathrm{I}} n_1 + c_{12}^{\mathrm{I}} n_0 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}}(n_1 + n_0)m_1 + c_2^{\mathrm{II}}(n_2 + n_0)m_2 - C_0\big] = 0, \tag{6.11}$$

$$\lambda_0 n_0 = 0, \quad \lambda_1 n_1 = 0, \quad \lambda_2 n_2 = 0 \tag{6.12}$$

To solve for λ, let us take the derivatives w.r.t. m_1, m_2 in (6.10):

$$\frac{\partial L}{\partial m_1} = -\frac{1}{Nm_1^2}\bar{S}_{1w}^2 - \lambda c_1^{\mathrm{II}}(n_1 + n_0) = 0, \tag{6.13}$$

$$\frac{\partial L}{\partial m_2} = -\frac{1}{Nm_2^2}\bar{S}_{2w}^2 - \lambda c_2^{\mathrm{II}}(n_1 + n_0) = 0, \tag{6.14}$$

so

$$-\lambda = \frac{\bar{S}_{1w}^2}{Nm_1^2 c_1^{\mathrm{II}}(n_1 + n_0)} = \frac{\bar{S}_{2w}^2}{Nm_2^2 c_2^{\mathrm{II}}(n_2 + n_0)} > 0 \tag{6.15}$$

as all terms in the right hand side are positive, provided there is any variation in at least one cluster at each point in time. By (6.11), strict inequality for λ means that the restriction is binding, or active, and the budget of the survey is fully used.

Next, let us take the derivatives w.r.t. n_1, n_2, n_0:

$$\frac{\partial L}{\partial n_1} = -\frac{S_{1b}^2}{M^2(n_1 + n_0)^2} + \frac{2n_0\rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1 + n_0)^2(n_2 + n_0)M^2} + \lambda_1 - \lambda(c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1) = 0, \tag{6.16}$$

$$\frac{\partial L}{\partial n_2} = -\frac{S_{2b}^2}{M^2(n_2 + n_0)^2} + \frac{2n_0\rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1 + n_0)(n_2 + n_0)^2M^2} + \lambda_2 - \lambda(c_2^{\mathrm{I}} + c_2^{\mathrm{II}} m_1) = 0, \tag{6.17}$$

$$-\lambda_1 = -\frac{S_{1b}^2}{M^2(n_1+n_0)^2} + \frac{2n_0\rho^{\mathrm{I}}S_{1b}S_{2b}}{(n_1+n_0)^2(n_2+n_0)M^2} + \frac{\bar{S}_{1w}^2(c_1^{\mathrm{I}}+c_1^{\mathrm{II}}m_1)}{Nm_1^2c_1^{\mathrm{II}}(n_1+n_0)}, \tag{6.18}$$

$$-\lambda_2 = -\frac{S_{2b}^2}{M^2(n_2+n_0)^2} + \frac{2n_0\rho^{\mathrm{I}}S_{1b}S_{2b}}{(n_1+n_0)(n_2+n_0)^2M^2} + \frac{\bar{S}_{2w}^2(c_2^{\mathrm{I}}+c_2^{\mathrm{II}}m_2)}{Nm_2^2c_2^{\mathrm{II}}(n_2+n_0)}, \tag{6.19}$$

$$\begin{aligned}\frac{\partial L}{\partial n_0} = -\frac{S_{1b}^2}{M^2(n_1+n_0)^2} - \frac{S_{2b}^2}{M^2(n_2+n_0)^2} + \lambda_0 - \lambda(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}}m_1 + c_2^{\mathrm{II}}m_2) - \\ -\frac{2\rho^{\mathrm{I}}S_{1b}S_{2b}(n_1n_2-n_0^2)}{(n_1+n_0)^2(n_2+n_0)^2M^2} = 0,\end{aligned} \tag{6.20}$$

$$\begin{aligned}-\lambda_0 = -\frac{S_{1b}^2}{M^2(n_1+n_0)^2} - \frac{S_{2b}^2}{M^2(n_2+n_0)^2} - \frac{2\rho^{\mathrm{I}}S_{1b}S_{2b}(n_1n_2-n_0^2)}{(n_1+n_0)^2(n_2+n_0)^2M^2} + \\ + \frac{\bar{S}_{1w}\bar{S}_{2w}}{Nm_1m_2(c_1^{\mathrm{II}}c_2^{\mathrm{II}}(n_1+n_0)(n_2+n_0))^{\frac{1}{2}}}(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}}m_1 + c_2^{\mathrm{II}}m_2) = 0\end{aligned} \tag{6.21}$$

The signs of λ_1, λ_2, λ_0 cannot be determined, as the relative sizes of terms in (6.18), (6.19), (6.21) vary, in general. Thus the only feasible way to solve (6.8) is to go over conditions (6.12), find solutions for specific cases, and compare them.

Taking into account the symmetry with respect to time and the fact that $(n_0+n_1)(n_0+n_2) > 0$ (which means that a positive number of clusters are indeed sampled, either repeatedly (n_0) or independently (n_1, n_2)), the following three cases are 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.

6.1 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. (6.18) and (6.19) then give

$$0 = -\frac{S_{1b}^2}{M^2n_1^2} + \frac{\bar{S}_{1w}^2(c_1^{\mathrm{I}}+c_1^{\mathrm{II}}m_1)}{Nm_1^2c_1^{\mathrm{II}}n_1}, \quad n_1 = \frac{S_{1b}^2Nm_1^2c_1^{\mathrm{II}}}{M^2\bar{S}_{1w}^2(c_1^{\mathrm{I}}+c_1^{\mathrm{II}}m_1)} \tag{6.22}$$

$$0 = -\frac{S_{2b}^2}{M^2n_2^2} + \frac{\bar{S}_{2w}^2(c_2^{\mathrm{I}}+c_2^{\mathrm{II}}m_2)}{Nm_2^2c_2^{\mathrm{II}}n_2}, \quad n_2 = \frac{S_{2b}^2Nm_2^2c_2^{\mathrm{II}}}{M^2\bar{S}_{2w}^2(c_2^{\mathrm{I}}+c_2^{\mathrm{II}}m_2)} \tag{6.23}$$

The other two equations to close the system are the total cost equation (6.6) and its Lagrange multiplier (6.15), which can be written as

$$c_1^{\mathrm{I}} n_1 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}} n_1 m_1 + c_2^{\mathrm{II}} n_2 m_2 = C_0, \tag{6.24}$$

$$\frac{\bar{S}_{1w}^2}{m_1^2 c_1^{\mathrm{II}} n_1} = \frac{\bar{S}_{2w}^2}{m_2^2 c_2^{\mathrm{II}} n_2} \tag{6.25}$$

Combining (6.22), (6.23) and (6.24), we obtain

$$\frac{N}{M^2}\left(\frac{S_{1b}^2 m_1^2 c_1^{\mathrm{II}}}{\bar{S}_{1w}^2} + \frac{S_{2b}^2 m_2^2 c_2^{\mathrm{II}}}{\bar{S}_{2w}^2}\right) = C_0 \tag{6.26}$$

which gives the total "scale" of the survey.

To solve for the parameters of the design, let us first assume that the costs and variances do not change between the two periods:

$$S_{1b}^2 = S_{2b}^2 = S_b^2,\ \bar{S}_{1w}^2 = \bar{S}_{2w}^2 = \bar{S}_w^2,\ c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = c^{\mathrm{I}},\ c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = c^{\mathrm{II}} \tag{6.27}$$

Then $m_1 = m_2 = m$ is found from

$$\frac{2N}{M^2}\frac{S_b^2 m^2 c^{\mathrm{II}}}{\bar{S}_w^2} = C_0, \quad m = M\sqrt{\frac{C_0 \bar{S}_w^2}{2Nc^{\mathrm{II}} S_b^2}} \tag{6.28}$$

Then also

$$n = \frac{C_0}{2(c^{\mathrm{I}} + c^{\mathrm{II}} m)} = \frac{C_0}{2\left[c^{\mathrm{I}} + M\sqrt{C_0 c^{\mathrm{II}} \bar{S}_w^2 / 2NS_b^2}\,\right]} \tag{6.29}$$

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

The variance of the difference estimator given by (6.3) is

$$\begin{aligned} \mathbb{V}_{e,i}[d] &= 2\frac{N-n}{nNM^2}S_{1b}^2 + 2\frac{1}{NM}\frac{M-m}{m}\bar{S}_w^2 - 2\frac{\rho^{\mathrm{I}} S_b^2}{nM^2} \approx \\ &\approx \frac{2S_b^2}{nM^2} + \frac{2\bar{S}_w^2}{Nm} - \frac{2\rho^{\mathrm{I}} S_b^2}{nM^2} = \\ &= \frac{4S_b^2\left[c^{\mathrm{I}} + \sqrt{M^2 C_0 c^{\mathrm{II}} \bar{S}_w^2 / 2NS_b^2}\,\right]}{C_0 M^2}(1-\rho^{\mathrm{I}}) + 2\sqrt{\frac{2c^{\mathrm{II}} S_b^2 \bar{S}_w^2}{NC_0 M^2}} \end{aligned} \tag{6.30}$$

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".

Let us go back to the situation when the population parameters and costs change over time. Rephrasing (6.26) gives

$$\frac{NS_{1b}^2 c_1^{\mathrm{II}}}{M^2 S_{1w}^2} m_1^2 \left(1 + \frac{S_{2b}^2 \bar{S}_{1w}^2 c_2^{\mathrm{II}}}{S_{1b}^2 \bar{S}_{2w}^2 c_1^{\mathrm{II}}}\frac{m_2^2}{m_1^2}\right) = C_0 \tag{6.31}$$

Denoting

$$\begin{aligned} A_1 &= \frac{M^2 \bar{S}_{1w}^2 c_1^{\mathrm{I}}}{S_{1b}^2 N c_1^{\mathrm{II}}}, \quad B_1 = \frac{M^2 \bar{S}_{1w}^2}{S_{1b}^2 N}, \quad C_1 = \frac{\bar{S}_{1w}^2}{c_1^{\mathrm{II}}} \\ A_2 &= \frac{M^2 \bar{S}_{2w}^2 c_2^{\mathrm{I}}}{S_{2b}^2 N c_2^{\mathrm{II}}}, \quad B_2 = \frac{M^2 \bar{S}_{2w}^2}{S_{2b}^2 N}, \quad C_2 = \frac{\bar{S}_{2w}^2}{c_2^{\mathrm{II}}} \end{aligned} \tag{6.32}$$

we get

$$n_1 = \frac{m_1^2}{A_1 + B_1 m_1}, \quad n_2 = \frac{m_2^2}{A_2 + B_2 m_2}, \quad \frac{C_1}{m_1^2 n_1} = \frac{C_2}{m_2^2 n_2}, \tag{6.33}$$

or

$$\frac{C_1(A_1 + B_1 m_1)}{m_1^4} = \frac{C_2(A_2 + B_2 m_2)}{m_2^4} \tag{6.34}$$

This is a quartic equation on m_2 with no explicit closed form solution. If

$$A_2 = A_1 + \delta_A, \;\; B_2 = B_1 + \delta_B, \;\; C_2 = C_1 + \delta_C, \;\; m_2 = m_1(1 + \delta_m), \tag{6.35}$$

where δ_A, δ_B, δ_C depend on the changes in costs and population variances between the two rounds of the survey, and $\delta_Z \ll Z$ denotes a variation in Z of small magnitude, then to the first order in δ's,

$$\begin{aligned} \delta_m &= \frac{(A_1 + B_1 m_1)\delta_C + C_1 \delta_A + C_1 m_1 \delta_B}{C_1(4A_1 + 3B_1 m_1)} = \\ &= \frac{c_1^{\mathrm{II}}}{M^2 \bar{S}_{1w}^2} \frac{M^2(c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)\delta_C + S_{1b}^2 N(\delta_A + m_1 \delta_B)}{4c_1^{\mathrm{I}} + 3c_1^{\mathrm{II}} m_1} \end{aligned} \tag{6.36}$$

which can be plugged to (6.31), using (6.28) for m_1. Taking everything together gives

$$m_1 = \sqrt{\frac{M^2 C_0 \bar{S}_{1w}^2}{(2 + \Delta) N c_1^{\mathrm{II}} S_{1b}^2}} \approx m\big(1 - \frac{\Delta}{4}\big), \tag{6.37}$$

$$\begin{aligned} \Delta = \frac{S_{2b}^2 \bar{S}_{1w}^2 c_2^{\mathrm{II}}}{S_{1b}^2 \bar{S}_{2w}^2 c_1^{\mathrm{II}}} - 1 + \delta_m &= \frac{S_{2b}^2 \bar{S}_{1w}^2 c_2^{\mathrm{II}}}{S_{1b}^2 \bar{S}_{2w}^2 c_1^{\mathrm{II}}} - 1 + \\ + \frac{c_1^{\mathrm{II}}}{M^2 \bar{S}_{1w}^2 (4c_1^{\mathrm{I}} + 3c_1^{\mathrm{II}} m)} &\left\{ M^2(c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m) \left(\frac{\bar{S}_{2w}^2}{c_2^{\mathrm{II}}} - \frac{\bar{S}_{1w}^2}{c_1^{\mathrm{II}}} \right) + \right. \\ \left. + S_{1b}^2 N \left[\frac{M^2 \bar{S}_{2w}^2 c_2^{\mathrm{I}}}{S_{2b}^2 N c_2^{\mathrm{II}}} - \frac{M^2 \bar{S}_{1w}^2 c_1^{\mathrm{I}}}{S_{1b}^2 N c_1^{\mathrm{II}}} + m \left(\frac{M^2 \bar{S}_{2w}^2}{S_{2b}^2 N} - \frac{M^2 \bar{S}_{1w}^2}{S_{1b}^2 N} \right) \right] \right\}&, \end{aligned} \tag{6.38}$$

$$m_2 = m_1(1 + \delta_m) \approx m\big(1 - \frac{\Delta}{4}\big)(1 + \delta_m) \tag{6.39}$$

and further n_1, n_2 found from (6.22) and (6.23).

6.2 Cluster-panel design

Let us now study the design at the opposite end of the spectrum. If the design with $n_1 = n_2 = 0$ is optimal, then from KKT conditions (6.12), $\lambda_0 = 0$. With those conditions, the system of equations to find the parameters of the optimal design is

$$\frac{\bar{S}_{1w}^2}{m_1^2 c_1^{\mathrm{II}}} = \frac{\bar{S}_{2w}^2}{m_2^2 c_2^{\mathrm{II}}}, \tag{6.40}$$

$$c_{12}^{\mathrm{I}} n_0 + c_1^{\mathrm{II}} n_0 m_1 + c_2^{\mathrm{II}} n_0 m_2 = C_0, \tag{6.41}$$

$$-\frac{S_{1b}^2}{M^2 n_0} - \frac{S_{2b}^2}{M^2 n_0} + \frac{2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n_0 M^2} + \frac{\bar{S}_{1w}\bar{S}_{2w}}{N m_1 m_2 (c_1^{\mathrm{II}} c_2^{\mathrm{II}})^{\frac{1}{2}}} (c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2) = 0 \tag{6.42}$$

Then

$$\begin{aligned}
m_2 &= m_1 \sqrt{\frac{\bar{S}_{2w}^2 c_1^{\mathrm{II}}}{\bar{S}_{1w}^2 c_2^{\mathrm{II}}}} \equiv \kappa m_1, \\
n_0 &= \frac{(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}) N \kappa m_1^2 (c_1^{\mathrm{II}} c_2^{\mathrm{II}})^{\frac{1}{2}}}{M^2 \bar{S}_{1w} \bar{S}_{2w} \big(c_{12}^{\mathrm{I}} + m_1 (c_1^{\mathrm{II}} + \kappa c_2^{\mathrm{II}})\big)} \equiv \frac{U m_1^2}{\big(c_{12}^{\mathrm{I}} + m_1 (c_1^{\mathrm{II}} + \kappa c_2^{\mathrm{II}})\big)}, \\
U &= \frac{(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}) N \kappa (c_1^{\mathrm{II}} c_2^{\mathrm{II}})^{\frac{1}{2}}}{M^2 \bar{S}_{1w} \bar{S}_{2w}}, \\
U m_1^2 = C_0, &\quad m_1 = \sqrt{C_0/U}, \quad m_2 = \kappa \sqrt{C_0/U}, \quad n_0 = \frac{C_0}{c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2}
\end{aligned} \tag{6.43}$$

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 (6.3) as

$$\begin{aligned}
\mathbb{V}[d] &= \frac{N - n_0}{n_0 N M^2} S_{1b}^2 + \frac{1}{NM} \frac{M - m_1}{m_1} \bar{S}_{1w}^2 + \\
&+ \frac{N - n_0}{n_0 N M^2} S_{2b}^2 + \frac{1}{NM} \frac{M - m_2}{m_2} \bar{S}_{2w}^2 - 2 \frac{\rho^{\mathrm{I}} S_{1b} S_{2b}}{n_0 M^2} \approx \\
&\approx \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n_0 M^2} + \frac{1}{N} \Big(\frac{\bar{S}_{1w}^2}{m_1} + \frac{\bar{S}_{2w}^2}{m_2} \Big) = \\
&= \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{C_0 M^2} \Big(c_{12}^{\mathrm{I}} + \frac{C_0^{\frac{1}{2}}}{U^{\frac{1}{2}}} (c_1^{\mathrm{II}} + \kappa c_2^{\mathrm{II}}) \Big) + \frac{U^{\frac{1}{2}}}{N C_0^{\frac{1}{2}}} \big(\bar{S}_{1w}^2 + \bar{S}_{2w}^2 \big)
\end{aligned} \tag{6.44}$$

Let us now invoke the assumption of no changes in population variances and prices between two time periods. Then

$$S^2_{1b}=S^2_{2b}=S^2_b,\ \bar S^2_{1w}=\bar S^2_{2w}=\bar S^2_w,\ c^{\rm II}_1=c^{\rm II}_2=c^{\rm II},\ \kappa=1,$$
$$U=\frac{2S^2_b(1-\rho^{\rm I})Nc^{\rm II}}{M^2\bar S^2_w}, \tag{6.45}$$

and the variance (6.44) becomes

$$\mathbb{V}_{e,p}[d]=\frac{2S^2_b(1-\rho^{\rm I})}{C_0M^2}\Bigl(c^{\rm I}_{12}+2\frac{{C_0}^{\frac12}}{U^{\frac12}}c^{\rm II}\Bigr)+2\frac{U^{\frac12}}{N{C_0}^{\frac12}}\bar S^2_w=$$
$$=\frac{2S^2_b(1-\rho^{\rm I})}{C_0M^2}\left(c^{\rm I}_{12}+2\sqrt{\frac{M^2\bar S^2_wC_0c^{\rm II}}{2S^2_b(1-\rho^{\rm I})N}}\right)+2\sqrt{\frac{2S^2_b\bar S^2_w(1-\rho^{\rm I})c^{\rm II}}{M^2NC_0}} \tag{6.46}$$

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

6.3 Comparison of the independent and panel-cluster designs

We now can compare (6.30) and (6.46). The difference of two variances is

$$\mathbb{V}_{e,p}[d]-\mathbb{V}_{e,i}[d]=$$
$$=\frac{2S^2_b(1-\rho^{\rm I})}{C_0M^2}\left[c^{\rm I}_{12}-2c^{\rm I}+2\sqrt{\frac{M^2\bar S^2_wC_0c^{\rm II}}{2S^2_bN}}\left(\frac{1}{\sqrt{1-\rho^{\rm I}}}-1\right)\right]+$$
$$+2\sqrt{\frac{2S^2_b\bar S^2_wc^{\rm II}}{M^2NC_0}}(\sqrt{1-\rho^{\rm I}}-1)=$$
$$=\frac{2S^2_b(1-\rho^{\rm I})}{C_0M^2}(c^{\rm I}_{12}-2c^{\rm I})-2\sqrt{\frac{2\bar S^2_wS^2_bc^{\rm II}}{M^2NC_0}}\bigl(1-\sqrt{1-\rho^{\rm I}}\bigr)^2 \tag{6.47}$$

The last term is always negative, and the cluster-panel design is guaranteed to be more efficient when $c^{\rm I}_{12}<2c^{\rm I}$, i.e., when revising the clusters indeed provides cost savings. Note also that as $\rho^{\rm I}\to 1$ (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:

$$c^{\rm I}_{12}<2c^{\rm I}+\sqrt{\frac{2\bar S^2_wM^2c^{\rm II}C_0}{S^2_bN(1-\rho^{\rm I})^2}}\bigl(1-\sqrt{1-\rho^{\rm I}}\bigr)^2 \tag{6.48}$$

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

6.4 Intermediate case

In this section, we consider a design that has all of $n_0, n_1, n_2 > 0$. For this design to be optimal, the corresponding Lagrange multipliers have to be zero:

$$\lambda_0 = \lambda_1 = \lambda_2 = 0 \tag{6.49}$$

Let us consider the implications for the first order conditions (6.18)–(6.21):

$$0 = -\frac{S_{1b}^2}{M^2(n_1+n_0)} + \frac{2n_0\rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1+n_0)(n_2+n_0)M^2} + \frac{\bar{S}_{1w}^2 (c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)}{N m_1^2 c_1^{\mathrm{II}}}, \tag{6.50}$$

$$0 = -\frac{S_{2b}^2}{M^2(n_2+n_0)} + \frac{2n_0\rho^{\mathrm{I}} S_{1b} S_{2b}}{(n_1+n_0)(n_2+n_0)M^2} + \frac{\bar{S}_{2w}^2 (c_2^{\mathrm{I}} + c_2^{\mathrm{II}} m_2)}{N m_2^2 c_2^{\mathrm{II}}}, \tag{6.51}$$

$$\begin{aligned} 0 = &-\frac{S_{1b}^2}{M^2(n_1+n_0)^2} - \frac{S_{2b}^2}{M^2(n_2+n_0)^2} - \frac{2\rho^{\mathrm{I}} S_{1b} S_{2b}(n_1 n_2 - n_0^2)}{(n_1+n_0)^2(n_2+n_0)^2 M^2} + \\ &+ \frac{\bar{S}_{1w}\bar{S}_{2w}}{N m_1 m_2 (c_1^{\mathrm{II}} c_2^{\mathrm{II}} (n_1+n_0)(n_2+n_0))^{\frac{1}{2}}} (c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2) = 0 \end{aligned} \tag{6.52}$$

The remaining equations to close the system are (6.6) and (6.15):

$$c_1^{\mathrm{I}} n_1 + c_{12}^{\mathrm{I}} n_0 + c_2^{\mathrm{I}} n_2 + c_1^{\mathrm{II}}(n_1+n_0)m_1 + c_2^{\mathrm{II}}(n_2+n_0)m_2 = C_0, \tag{6.53}$$

$$\frac{\bar{S}_{1w}^2}{m_1^2 c_1^{\mathrm{II}}(n_1+n_0)} = \frac{\bar{S}_{2w}^2}{m_2^2 c_2^{\mathrm{II}}(n_2+n_0)} \tag{6.54}$$

As the problem appears intractable, let us impose the equality of the two periods restrictions (6.27). Then by symmetry of the problem, $n_1 = n_2 \equiv n$, $m_1 = m_2 \equiv m$, and

$$\frac{S_b^2(n + n_0 - 2n_0\rho^{\mathrm{I}})}{M^2(n+n_0)^2} = \frac{\bar{S}_w^2(c^{\mathrm{I}} + c^{\mathrm{II}} m)}{N m^2 c^{\mathrm{II}}}, \tag{6.55}$$

$$\frac{2S_b^2(n + n_0 + \rho^{\mathrm{I}}(n - n_0))}{M^2(n+n_0)^2} = \frac{\bar{S}_w^2}{N m^2 c^{\mathrm{II}}}(c_{12}^{\mathrm{I}} + 2c^{\mathrm{II}} m) \tag{6.56}$$

$$2c^{\mathrm{I}} n + c_{12}^{\mathrm{I}} n_0 + 2c^{\mathrm{II}}(n+n_0)m = C_0 \tag{6.57}$$

Introducing

$$\nu = \frac{n_0}{n}, \quad \pi^{-1} = 1 + \nu^{-1} \tag{6.58}$$

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

$$\begin{gathered} m = \frac{c^{\mathrm{I}}(1 + \nu + \rho^{\mathrm{I}} - \rho^{\mathrm{I}}\nu) - c_{12}^{\mathrm{I}}(1 + \nu - 2\rho^{\mathrm{I}}\nu)}{c^{\mathrm{II}}(1 + \nu - 3\rho^{\mathrm{I}}\nu - \rho^{\mathrm{I}})} \equiv \frac{A + B\nu}{C + D\nu}, \\ A = c^{\mathrm{I}}(1+\rho^{\mathrm{I}}) - c_{12}^{\mathrm{I}}, \quad B = c^{\mathrm{I}}(1-\rho^{\mathrm{I}}) - c_{12}^{\mathrm{I}}(1 - 2\rho^{\mathrm{I}}), \\ C = c^{\mathrm{II}}(1-\rho^{\mathrm{I}}), \quad D = c^{\mathrm{II}}(1 - 3\rho^{\mathrm{I}}) \end{gathered} \tag{6.59}$$

The coefficients A, B, C and D as functions of ρ^{I} are shown on Fig. 1 (a).

For m to be positive, $(A + B\nu)(C + D\nu) > 0$. The zeroes of the latter polynomial are $-A/B$ and $-C/D$. Let us study those expressions as functions of ρ^{I} to identify the regions where m can be positive.

For $\rho^{\mathrm{I}} = 0$, $A = B = c^{\mathrm{I}} - c^{\mathrm{I}}_{12} < 0$, so $-A/B = -1$. Also, $C = D = c^{\mathrm{II}}$, so $-C/D = -1$ as well. In this case $m = A/C = (c^{\mathrm{I}} - c^{\mathrm{I}}_{12})/c^{\mathrm{II}} < 0$, so the mixed design cannot be optimal.

As ρ^{I} increases, A and B increase, while C and D decrease. By continuity, the "optimal" m continues to be negative in the neighborhood of zero, thus the mixed design cannot be optimal near $\rho^{\mathrm{I}} = 0$. The sign of m changes when $B(\rho^{\mathrm{I}})$ crosses zero. It occurs when

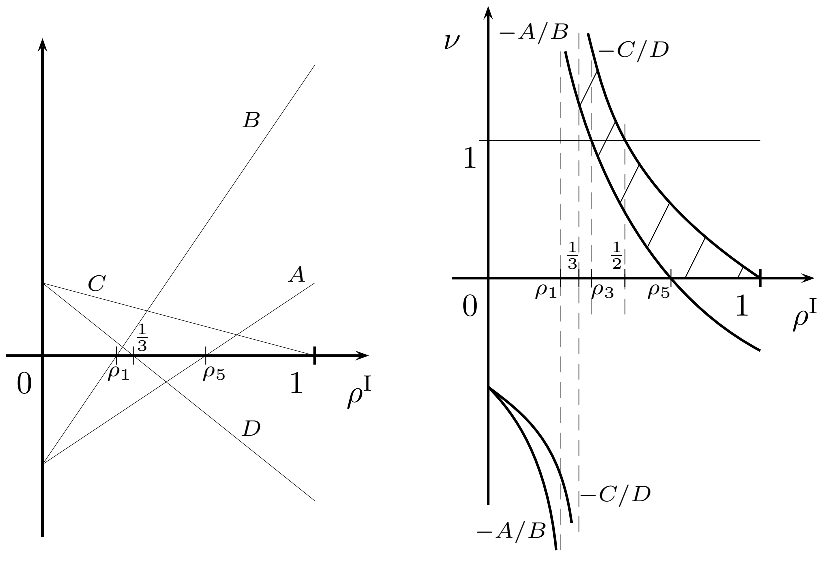
$$\rho = \rho_1 \equiv \frac{c^{\mathrm{I}}_{12} - c^{\mathrm{I}}}{2c^{\mathrm{I}}_{12} - c^{\mathrm{I}}} \tag{6.60}$$

If we assume

$$c^{\mathrm{I}} < c^{\mathrm{I}}_{12} < 2c^{\mathrm{I}} \tag{6.61}$$

(i.e., the cost of revisiting the cluster used before is greater than 0, but less than the cost of sampling the new cluster), then it can be shown that $\rho_1 < 1/3 \equiv \rho_2$, the point where the line D crosses zero. Thus for $\rho_1 < \rho \leq \rho_2$, both $C + D\nu$ is positive, and $A + B\nu$ is positive for large enough ν, namely, $\nu \geq -A/B$.

As ρ^{I} grows further, D crosses zero at $\rho = \rho_2$, and the root $-C/D$ flips its sign. Now, $C + D\nu$ is only positive for small enough ν, namely, $\nu < -C/D$.



(a) (b)

Figure 1: (a) Coefficients A, B, C, D as functions of ρ; (b) Critical points of ν and feasible regions.

Hence the feasible values of ν are in the range $-A/B \le \nu < -C/D$.

A pictorial representation of $-A/B$ and $-C/D$ as functions of ρ^{I} are given on Fig. 1 (b). The shaded region is the range of ν's for which m can be positive, according to (6.59).

Two more characteristic values of ρ^{I} may be of interest that correspond to $\nu = 1$ (i.e., $n_0 = n_1 = n_2$, and the sample is split 50-50 between the independent and panel components). The lower bound of correlations at which this can happen is given by $A = -B$, which is solved by

$$\rho_3 \equiv \frac{c_{12}^{\mathrm{I}} - c^{\mathrm{I}}}{c_{12}^{\mathrm{I}}}, \quad \rho_1 < \rho_3 < \frac{1}{2} \tag{6.62}$$

The upper bound is given by the solution to $C = -D$, and that is $\rho_4 = 1/2$. The line corresponding to $\nu = 1$ (i.e. the sample is split 50-50 between the new and repeated clusters) is also shown on Fig. 1 (b).

Finally, the last characteristic value of ρ^{I} is the one where A changes its sign:

$$\rho_5 \equiv \frac{c_{12}^{\mathrm{I}} - c^{\mathrm{I}}}{c^{\mathrm{I}}} = \rho_3 \frac{c_{12}^{\mathrm{I}}}{c^{\mathrm{I}}}, \quad \rho_3 < \rho_5 < 1 \tag{6.63}$$

The inequalities in (6.62) and (6.63) follow from (6.61).

Let us go back to (6.55)–(6.57). Subtracting (6.56) from (6.55), one gets

$$\begin{aligned} \frac{\bar{S}_w^2 (c_{12}^1 - c^{\mathrm{I}} + c^{\mathrm{II}} m)}{N m^2} &= \frac{2 S_b^2}{M^2 (n + n_0)}, \\ n = m^2 \frac{2 S_b^2 N}{\bar{S}_w^2 M^2 (1+\nu)(c_{12}^1 - c^{\mathrm{I}} + c^{\mathrm{II}} m)} &= \\ = \frac{(A + B\nu)^2}{C + D\nu} \frac{2 S_b^2 N}{\bar{S}_w^2 M^2 (1+\nu)((c_{12}^1 - c^{\mathrm{I}})(C + D\nu) + c^{\mathrm{II}}(A + B\nu))} & \end{aligned} \tag{6.64}$$

Plugging this to (6.57) gives

$$C_0 \frac{\bar{S}_w^2 M^2}{2 S_b^2 N} (1+\nu) = \frac{(A + B\nu)^2}{(C + D\nu)^2} \frac{(2c^{\mathrm{I}} + c_{12}^{\mathrm{I}} \nu)(C + D\nu) + 2c^{\mathrm{II}}(1+\nu)(A + B\nu)}{(c_{12}^1 - c^{\mathrm{I}})(C + D\nu) + c^{\mathrm{II}}(A + B\nu)} \tag{6.65}$$

Denoting the constant in front of $1 + \nu$ in the LHS by Q, one gets the following equation for ν which is a fourth order polynomial in ν:

$$\begin{aligned} P(\nu; \rho^{\mathrm{I}}) \equiv Q(1+\nu)(C + D\nu)^2 \big[(c_{12}^1 - c^{\mathrm{I}})(C + D\nu) + c^{\mathrm{II}}(A + B\nu)\big] - \\ -(A + B\nu)^2 \big[(2c^{\mathrm{I}} + c_{12}^{\mathrm{I}} \nu)(C + D\nu) + 2c^{\mathrm{II}}(1+\nu)(A + B\nu)\big] = 0 \end{aligned} \tag{6.66}$$

It is easy to establish that a solution exists in the feasible domain of Fig. 1 (b). This domain corresponds to values of ν such that both $A + B\nu$ and $C + D\nu$ are positive. Hence, both terms in the square brackets are positive. Further, $P(-A/B; \rho^{\mathrm{I}}) = Q(1+\nu)(C + D\nu)^3 (c_{12}^{\mathrm{I}} - c^{\mathrm{I}}) > 0$ whenever $C + D\nu$ is positive.

For $\rho_1 < \rho^{\mathrm{I}} < 1/3$, let us look at the behavior at infinity: $P(+\infty; \rho^{\mathrm{I}}) \sim \big\{QD^2\big[(c^{\mathrm{I}}_{12} - c^{\mathrm{I}})D + c^{\mathrm{II}}B\big] - B^2\big[c^{\mathrm{I}}_{12}D + 2c^{\mathrm{II}}B\big]\big\}\nu^4$. For ρ^{I} close to ρ_1, B is small, and $P(+\infty; \rho_1 + 0)$ can be further reduced to $QD^2(c^{\mathrm{I}}_{12} - c^{\mathrm{I}})D\nu^4 > 0$. For ρ^{I} close to $\rho_2 = 1/3$, D is small, and $P(+\infty; \rho_2 - 0)$ reduces to $-2c^{\mathrm{II}}B^3\nu^4 < 0$. Thus there is an intermediate ρ_6 such that $P(+\infty; \rho_6) = O(\nu^3)$ with the coefficient of ν^4 equal to zero. This also makes sense since for $\rho^{\mathrm{I}} = \rho_1 + 0$, the feasible regions have high n and m, which makes the design overly expensive.

For $\rho^{\mathrm{I}} > 1/3$, it is easy to see that $P(-C/D; \rho^{\mathrm{I}}) < 0$, and by continuity of $P(\cdot)$, there is a root between $-A/B$ and $-C/D$. The restriction $\nu > 0$ will also come to play for high values of the correlation: $P(0; \rho^{\mathrm{I}}) = QC^2\big[(c^1_{12} - c^{\mathrm{I}})C + c^{\mathrm{II}}A\big] - 2A^2(c^{\mathrm{I}}C + c^{\mathrm{II}}A)$ Clearly, $P(0; \rho_5) > 0$ as already established, and $P(0; 1) < 0$, so by continuity of $P(\cdot)$ with respect to ρ^{I} there exists an intermediate ρ_7 for which $P(0; \rho_7) = 0$.

Hence the following existence result is established:

$$\exists\, \rho_6 : \rho_1 < \rho_6 < \rho_2\ \exists\, \rho_7 : \rho_5 < \rho_7 < 1\ \forall \rho_6 < \rho^{\mathrm{I}} \le \rho_7\ \exists\, \nu : P(\nu; \rho^{\mathrm{I}}) = 0 \qquad (6.67)$$

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.

6.5 Numerical illustration

Without an analytical solution, it is impossible to compare the results of the analysis of the three designs. Let us then develop a simple numerical example to solve (6.66) for the optimal mixed design, and compare the results with the analytical cases of longitudinal and independent sampling designs.

Let us use the following set of parameters:

$$\begin{aligned} &N = 2000, \quad M = 200, \quad S_b = 1.5, \quad S_w = 1, \\ &c^{\mathrm{I}} = 1, \quad c^{\mathrm{I}}_{12} = 1.7, \quad c^{\mathrm{II}} = 0.25, \quad C_0 = 500, \end{aligned} \qquad (6.68)$$

In computing optimal designs, the practical restrictions were imposed:

$$\begin{gathered} 2 \le m \le M, \\ 2 \le n \le N \text{ for independent clusters design,} \\ 2 \le n_0 \le N \text{ for panel clusters design,} \\ 1 \le n, n_0 \le N \text{ for mixed design} \end{gathered} \qquad (6.69)$$

The action of those constraints produced some roughness in the graphs that follow, as well slight deviations in the budget (within a margin of 2 for the cluster-panel design, and with a wider range of fluctuations between 488 and 535 for the mixed design. The latter value is observed in the neighborhood of $\rho^{\mathrm{I}} = 1 - 0$ where the mixed design starts showing irregular behavior as it hits the constraints (6.69).

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 number of clusters sampled is shown on Fig. 2, and cluster sizes, on Fig. 3. The characteristic intertemporal correlations relevant for the mixed design are:

$$\rho_1 = 0.292,\ \rho_2 = 1/3,\ \rho_3 = 0.412,\ \rho_4 = 0.5,\ \rho_5 = 0.7,\ \rho_6 = 0.663,\ \rho_7 = 0.997$$

The horizontal solid line on Fig. 2 is the number of independent clusters for independent design; i.e., 10. The downward sloping solid line is the number of overlapping clusters in the cluster-panel design. As ρ^{I} increases, the 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 at least two clusters are to be taken to estimate the components of variance; and in fact $n_0 = 5$ clusters are sampled for $\rho^{\mathrm{I}} > 0.72$ as long as the optimal cluster size hits the restriction $m \leq M$. 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 cluster sizes on Fig. 3 show the differences between the designs. The panel cluster designs are equivalent to the independent cluster designs (except for slightly larger cluster sizes $m = 97$ due to savings offered by the cluster panel designs of the magnitude $2c^{\mathrm{I}} - c^{\mathrm{I}}_{12}$), but later try to get as most information from few clusters as possible as ρ^{I} approaches 1, and due to cluster size limits, $m = 197 \approx M$ for $\rho^{\mathrm{I}} > 0.72$. Interestingly, the cluster size remains fixed for the mixed design in the range between 32 and 34 for most of the correlation range. For $\rho^{\mathrm{I}} > 0.97$, the mixed design shows irregular behavior, as the optimal design requires $n_0 < 1$, which should have made it an independent cluster design that can no longer be called mixed.

The total sample sizes are approximately constant for the mixed designs (fluctuating between 896 and 918 for $\rho^{\mathrm{I}} < 0.4$, and between 891 and 896 for $0.4 < \rho^{\mathrm{I}} < 0.97$), and grow somewhat for the panel cluster designs (from 970 at $0 < \rho^{\mathrm{I}} < 0.14$, with $n_0 = 10$ clusters, to 985 at $\rho > 0.72$ with $n_0 = 5$ clusters).

Finally, the most important plot is that of the variance of the elementary estimate given on Fig. 4. 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.

6.6 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,

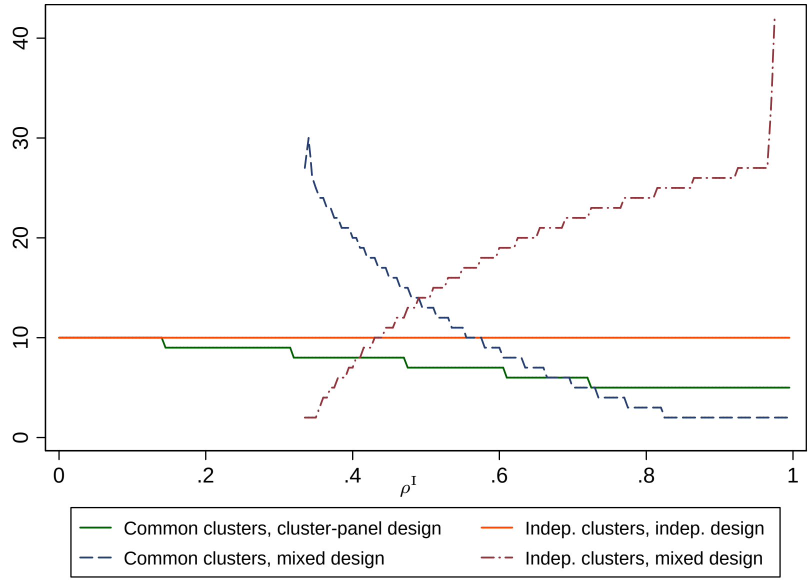


Figure 2: Number of clusters as a function of ρ^{I}.

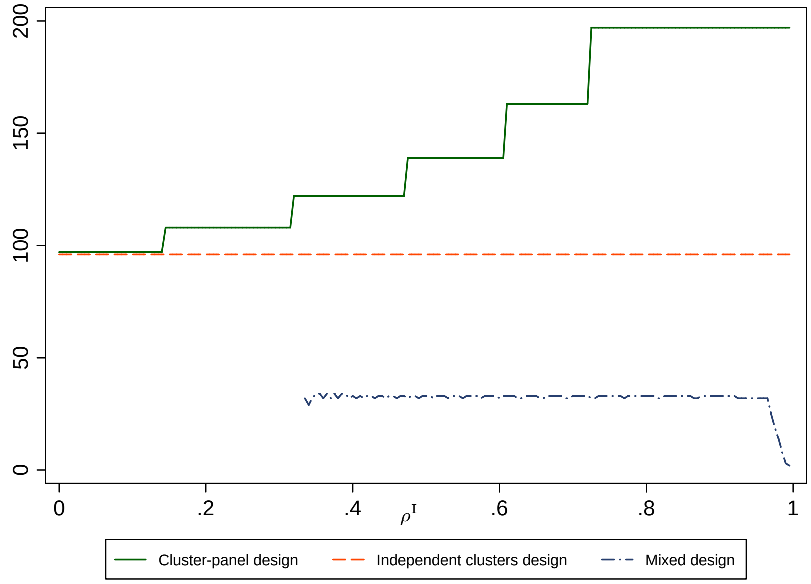


Figure 3: The size of a cluster m as a function of ρ^{I}.

and no new units appear The characteristics $Y_{ij}^{(1)}$, $Y_{ij}^{(2)}$ may be change, however. 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).

7 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

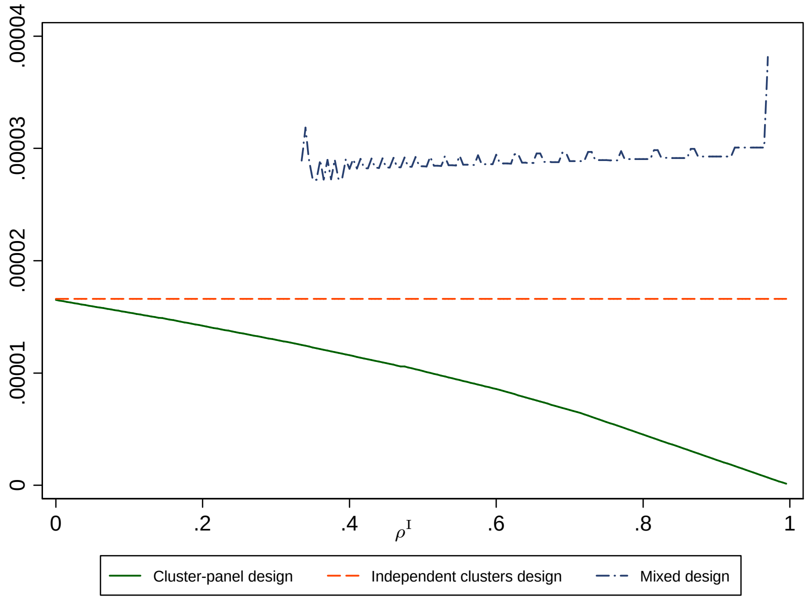


Figure 4: $\mathbb{V}\, d[\bar{y}^{(2)} - \bar{y}^{(1)}]$ as a function of ρ^{I}.

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

Item	Estimate	S.e.	DEFF
Contraceptive prevalence			
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
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
Longitudinal effect			0.976

Source: Bangladesh DHS, 1996 and 2000; authors' calculations.

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.

Appendix A gives an outline of Stata code used to specify the two alternative designs, and to estimate the difference of interest appropriately for those designs.

8 Simulation study

A small simulation was conducted to study the performance of the two estimators, Δ-naïve and Δ-design, as described in the previous section. The population consisted of 50 strata with means $\mu_h^{(1)} = \exp[0.01(h-1)]$, $h = 1, \dots, 50$ in the first period, so the means were shifting from 1.000 to 1.632. In the second period, there was growth in some strata: $\mu_h^{(1)} = \mu_h^{(1)}$ for the first 20 strata, $\mu_h^{(1)} = \mu_h^{(1)} + 0.25$ for the next 20 strata, and $\mu_h^{(1)} = \mu_h^{(1)} + 0.5$ for the last 10 strata, so that the total difference is $0 \cdot 0.4 + 0.25 \cdot 0.4 + 0.5 \cdot 0.2 = 0.2$. Each stratum consisted of 40 clusters of 500 units each. The variances within and between clusters, as well as the intertemporal correlations of the cluster means, were varying in the following ranges:

$$S_b \in \{0.1, 0.3, 1.0\}, \quad S_w \in \{0.2, 0.5, 1.5\}, \quad \rho^{\mathrm{I}} \in \{0, 0.25, 0.5, 0.75\}$$

The cluster means were sampled from a normal population, and then rescaled to have a finite population mean of 0 and variance of S_b^2. Likewise, the units were sampled from a normal population and rescaled to have a finite population mean of 0 and variance of S_w^2.

The number of clusters sampled varied between 2 and 20. For settings with 5 or less clusters sampled, each possible rearrangement into n_0, n_1 and n_2 was entertained. For settings with 10, 15 or 20 clusters, the number of common clusters n_0 and unique clusters $n_1 = n_2$ were varying in steps of 5. Finally, the number of units $m_1 = m_2 = m$ sampled from each cluster was 10, 20, or 50.

For each combinations of the settings described above, 39 Monte Carlo samples from the specific finite population were taken. This results in 126,360 total simulated samples for 3240 combinations of settings.

For each of those unique settings, the empirical variance of the difference estimates was found based on 39 simulated samples. The empirical design effect due to repeated use of the clusters (longitudinal effect) was defined as the ratio of the empirical variance of the difference estimate to the variance with the same parameters of the population (S_b, S_w, ρ^{I}), same sample size (n and m), but with zero overlap (independent sampling on each occasion).

The (asymptotic) functional form of the design effect is given by (5.15). Define

$$z = \frac{\rho\pi}{n} \tag{8.1}$$

The plot of the empirical DEFFs against z is given on Fig. 5. The agreement is not very strong; the R^2 in the linear regression is 0.13, and it goes up to

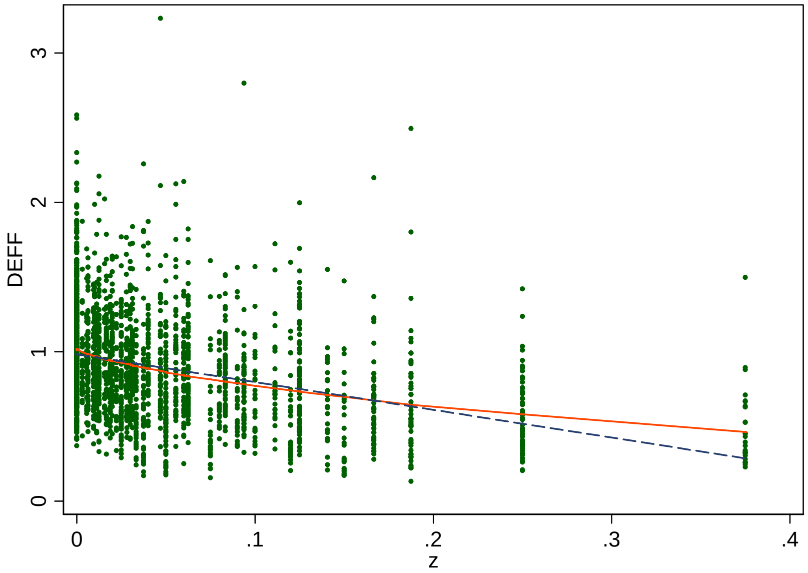


Figure 5: Hypothesized vs. empirical design effects. Solid line, lowess fit. Dotted line, prediction fit.

0.15 with a semiparametric (spline) regression. It is unclear what part of the regression error is due to simulation error, supposedly large due to small Monte Carlo sample for each setting (39 repetitions). The regression line shown as dotted is (s.e.s in parentheses, corrected for clustering on the same population)

$$DEFF = \underset{(0.0120)}{0.9824} - \underset{(0.2608)}{1.8572}z$$

The DEFF is not statistically significantly different from 1 for $z = 0$, and it does decrease linearly in the neighborhood of zero thus confirming the functional form of the DEFF behavior in finite samples.

The second part of the simulation consisted of taking samples from the finite population with $S_b = 0.2$, $S_w = 1$ and $\rho^{\mathrm{I}} = 0.5$. For this setting, the intraclass correlation (4.15) is low at about 0.04, mostly due to large $M = 500$. With the same sampling designs (n between 2 and 20, m between 10 and 50, etc.), 199 samples were taken for each combination that allowed for more accurate estimation of the variances.

The results reported below are for variances of the difference estimator $d[y^{(2)} - y^{(1)}]$, and various estimators of this variance. Table 2 reports the basic variances. The first column gives the sampling settings. The next four columns report the observed Monte Carlo standard deviation and the Monte Carlo means of the standard errors as reported by SRS design, Δ-naïve and Δ-design estimators of variance.

Table 2: Simulation results for estimator variances.

Sampling setting	MC s.d.	SRS s.e.	Δ-naïve	Δ-design
$m = 10$				
$n_0 = 0; n_1, n_2 = 2$	0.05073	0.04748	0.06945	0.06860
$n_0 = 2; n_1, n_2 = 0$	0.04985	0.04735	0.06967	0.04836
$n_0 = 0; n_1, n_2 = 20$	0.01471	0.01498	0.02078	0.01834
$n_0 = 20; n_1, n_2 = 0$	0.01403	0.01498	0.02080	0.01537
$m = 50$				
$n_0 = 0; n_1, n_2 = 2$	0.03522	0.02116	0.05602	0.05513
$n_0 = 2; n_1, n_2 = 0$	0.02677	0.02119	0.05721	0.02762
$n_0 = 0; n_1, n_2 = 20$	0.00852	0.00670	0.01645	0.01325
$n_0 = 20; n_1, n_2 = 0$	0.00745	0.00670	0.01646	0.00871

A few common features can be deducted from Table 2. The SRS estimator is almost always biased downwards, with the largest biases for high m and low n. This is typically a poor sampling design for a positive intraclass correlation, and will probably be avoided in practical situations. The Δ-design estimator is always conservative with a relative bias around 10 to 15% when the situation calls for the design appropriate estimator (i.e, $n_0 > 0$, $n_1 = n_2 = 0$). The

Table 3: Design effects.

Sampling setting	Reported DEFFs				Monte Carlo DEFFs	
	Total	IQR	Repeated	IQR	Total	Repeated
$m = 10$						
$n_0 = 0, n_1 = n_2 = 2$	2.0914	1.966–2.211	0.9760	0.959–0.997	1.3053	
$n_0 = 2, n_1 = n_2 = 0$	1.0547	0.888–1.196	0.4877	0.425–0.557	1.2294	0.9657
$n_0 = 0, n_1 = n_2 = 20$	1.4982	1.468–1.535	0.7789	0.764–0.793	0.9306	
$n_0 = 20, n_1 = n_2 = 0$	1.0532	1.024–1.080	0.5465	0.531–0.563	0.7692	0.9087
$m = 50$						
$n_0 = 0, n_1 = n_2 = 2$	6.7962	6.460–7.163	0.9686	0.953–0.987	7.6714	
$n_0 = 2, n_1 = n_2 = 0$	1.7165	1.465–1.934	0.2363	0.203–0.263	2.5483	0.5778
$n_0 = 0, n_1 = n_2 = 20$	3.9189	3.837–4.007	0.6501	0.637–0.661	2.6194	
$n_0 = 20, n_1 = n_2 = 0$	1.6921	1.639–1.746	0.2802	0.272–0.289	1.5347	0.7651

Δ-naïve estimator is always biased upwards, as it assumes more clustering than there is in the data. In fact, when $n_0 = 0$ and no clusters are sampled in common on purpose, the two estimators coincide in overly conservative performance. There still might be random overlap of clusters due to sampling, and hence Δ-design estimator is usually a little bit smaller than the Δ-naïve estimator.

Table 3 summarizes the design effects. The first four columns give the reported DEFF for the total effect (the ratio of SRS variance to the Δ-design variance estimates) and repeated clustering effect (the ratio of Δ-naive to Δ-design variance estimates). Note that those design effects also account for stratification. When $n_0 = 0$, the longitudinal/repeated cluster sampling design effect must be 1. Empirically, it is lower than 1 for the same reason of random overlap between clusters, and the difference from 1 is greater for sampling designs with larger number of sampled clusters $n_1 = n_2$ (20 vs. 2), and hence higher probability of the overlap.

The empirical, or Monte Carlo, design effects are defined here as follows: the total effect is the ratio of the Monte Carlo variance (column 2 of Table 2) to the SRS variance (column 3 of Table 2); the longitudinal design effect is the ratio of the empirical variances in Table 2 of the design with no overlap ($n_0 = 0$) to the empirical variance of the design with complete overlap ($n_0 = 2$ or $n_0 = 20$), i.e., the variance in the second row to the variance in the first row, the variance of the fourth row to the variance in the third row, etc. We see that all reported design effects are in fact quite misleading: none of the Monte Carlo DEFFs fall into the IQR of the reported DEFFs, neither for the total design effect, nor for the longitudinal design effect.

The evidence on behavior of the design effect with the number of clusters, predicted to be of the functional form of $1 - A\rho/n$, is mixed. With small number of observations per cluster ($m = 10$), the effect of increase in the number of clusters is to decrease the design effect moving it away from 1. With large number of observations per cluster ($m = 50$), which is more consistent with the asymptotic derivations outlined above, the design effect increases and gets closer to 1 as the number of sampled clusters increases. Again, the design effects here may be confounded with the stratification effect.

9 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, and π is the degree of overlap between two consecutive samples; and (ii) that for a given budget of the survey, the variance of the elementary estimator (difference in contemporaneous means) is smaller for the design that reuses clusters (referred to as cluster-panel designs) are more variance efficient for vs. the design where the samples are taken anew, with the difference in variances that 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 for helpful discussions and to William Kalsbeek for suggestions at the early stages of the paper. 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.

A 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: as the difference in means:

```
svymean depvar, by(year) deff
```

```
lincom depvar[2000] - depvar[1996], deff
```

or 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 estimatores of the variances of the difference.

References

Angeles, G., Guilkey, D. K. & Mroz, T. A. (1998), 'Purposive program placement and the estimation of family planning program effects in Tanzania', *Journal of the American Statistical Association* **93**(443), 884–899.

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust difference in differences estimates?', *Quarterly Journal of Economics* **119**(1), 249–275.

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

Di Pillo, G. & Palagi, L. (2002), Nonlinear programming, *in* P. M. Pardalos & M. G. C. Resende, eds, 'Handbook of Applied Optimization', Oxford University Press.

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. & Madow, W. G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons, New York.

Jahn, J. (1996), *Introduction to the Theory of Nonlinear Optimization*, 2nd revised edn, Springer, 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.

Korn, E. L. & Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley and Sons.

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.

Pitt, M., Rozenzweig, M. & Gibbons, D. (1993), 'The determinants and consequences of the placement of government programs in indonesia', *World Bank Economic Review* **7**(3), 319–348.

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

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.

Wolter, K. M. (1979), 'Composite estimation in finite population', *Journal of the American Statistical Association* **74**, 604–613.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

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